Genomic signatures of selection and non-adaptive evolution in a social microbe

Lima De Oliveira, Janaina

Award date:
2019

Awarding institution:
University of Bath

Genomic signatures of selection and non-adaptive evolution in a social microbe

Janaina Lima de Oliveira

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

September 2018

Janaina Lima de Oliveira

Table of contents

1 Introduction 1

1.1 A brief historical overview on the development of molecular evolution theories 1

1.2 Weak selection and synonymous variation 5

1.3 'Extended evolutionary null hypotheses' 6

1.4 Sociogenomics: addressing social evolution at the molecular level 10

1.5 Identifying dynamics of social evolution from molecular evolutionary signatures 12

1.6 *Dictyostelium discoideum* as a model system to understand evolutionary processes at the molecular level 16

1.7 Aims and approaches 18

2 Molecular evolution of social genes is consistent with signatures of the Red King process in a microbe 19

2.1 Abstract 20

2.2 Introduction 21

2.3 Results 24

2.3.1 Identification of social genes 24

2.3.2 Only sociality genes harbour increased sequence diversity 29

2.3.3 Elevated variation in sociality genes reflects weak selection 31

2.3.4 All classes of social genes primarily evolve under purifying selection 33

2.3.5 The RK process as a unifying explanation for patterns of molecular evolution 37

2.4 Discussion 39

2.5 Methods 45

2.5.1 Genomic DNA sequencing 45
2.5.2 Mapping and SNP calling 45
2.5.3 Intraspecific evolutionary statistics 46
2.5.4 Interspecific divergence 47
2.5.5 Identification of social genes 48
2.5.6 GO enrichment 51
2.5.7 Randomization procedure for significance testing 52
2.5.8 Data availability 52
2.6 Supplementary material 54
3 Processes shaping synonymous codon use in an extremely AT-biased genome 69
3.1 Abstract 70
3.2 Introduction 71
3.3 Results 73
3.3.1 AT-richer codons are used more frequently 74
3.3.2 Mutation bias explains a large proportion of synonymous codon use 75
3.3.3 CUB is also shaped by selection to optimize expression 79
3.4 Discussion 84
3.5 Methods 87
3.5.1 Synonymous codon frequencies and GC distribution across the genome 87
3.5.2 Nucleotide substitution matrix and GC_{eq} 87
3.5.3 Expected relative synonymous frequencies and identification of preferred codons 88
3.5.4 Parameters of translational and transcriptional selection 89

3.5.5 Expression levels and genes evolving under different selective constraints 91
3.6 Supplementary material 92
4 Amino acid composition is influenced by evolutionary processes shaping genomic and metabolic features in a microbe 94
4.1 Abstract 95
4.2 Methods 105
4.2.1 Amino acid frequencies 105
4.2.2 Base composition and metabolic cost parameters 105
4.2.3 Evolutionary tests 106
4.2.4 Gene expression 106
4.2.5 Statistical analyses 106
4.3 Supplementary material 107
5 General discussion 108
6 References 111

Acknowledgments

Firstly, I would like to thank my great team of supervisors led by Professor Jason Wolf, together with Dr Araxi Urrutia and Professor Christopher Thompson. As a group, your support has undoubtedly enriched this learning journey that we call a PhD. It is difficult to thank Jason enough without sounding hyperbolic. With his endless patience, he has taught me everything of the little I know about evolutionary biology, statistics, and social evolution, as well as the meaning of making good science. Araxi probably has to idea on how much she inspired me. As a Latin-American woman scientist, it is encouraging to feel represented by such a high-profile and friendly professional in this sphere. Chris, as well as people from his lab, has shared his great experience and knowledge (and data!), which certainly contributed to upgrade my work.

I would also like to thank Atahualpa, who is not listed as a supervisor for mere technicality. He has not only tutored me from my very first baby steps on programming language but accompanied me as a mentor and as a friend on my whole journey of a PhD-to-be. Former and current colleagues from Wolf lab have also technically and intellectually contributed to the development of this thesis: from the hard lab work of Emily, Bianca and Alex generating the very raw material data used throughout my work, to what I learned from Phil and Laurie about social evolution. Manu and Arodi, ephemeral members of the lab, shared their friendship and good moments over a beer or a coffee. My acknowledgement also extends to several people with whom I had the opportunity to discuss and improve the development of my project, as Professor Laurence Hurst, and my assessors Professors Ed Feil and Sam Sheppard.

I would also like to thank my family to the great support they have always provided me. Specially, my parents Nilza and Naldo, who always encouraged me to pursue my aspirations, my brother Tiago, who is always there (and here) for me, and Alisson, my beloved partner, whose loving heart – despite the distance – kept me warm in these foreign cold lands.

Last but not least, none of this work would have been possible without the financial support from the National Council for Scientific and Technological Development (CNPq).

Contributions

I confirm that the findings presented in this thesis are the result of my own work carried out with the advice and support of my supervisors Professor Jason Wolf, Dr Araxi Urrutia and Professor Christopher Thompson (University College of London), with the following exceptions.

In chapter 2 data collection for genomic sequencing was carried out by Emily Hind, Dr Bianca Reeksting and Dr Alexandria Holland. Gene expression experiment in clonal versus chimeric slugs was carried out by Dr Jennifer Engelmoer, and Dr Suzanne Battom Brown performed the gene expression experiment comparing prestalk and prespore cells.

Processing of genome and expression data used in chapters 2 to 4 where performed by Dr Atahualpa Castillo, with additional contributions from Dr Balint Stewart and Dr Nicole Gruenheit in chapter 2.

All aforementioned and Professor Laurence Hurst also contributed with valuable discussions and comments on the analyses and wording of the manuscripts here presented.

Results presented in chapter 2 have been submitted to the journal Nature Communications.

Results presented in chapters 3 and 4 are being prepared for submission.

Abstract

Evolutionary processes leave footprints across the genome. In its several forms, natural selection favours or removes mutations based on their fitness effects, which has consequences to patterns of standing variation, linkage disequilibrium, and rates of evolution. Genomic tools have allowed for a revolution in how we can study these problems, leading to great progress from the understanding of co-evolutionary dynamics in nature to the design of parasite-targeted drugs in medical sciences. However, despite the inarguable importance of selection, patterns across the genome are not necessarily a result of selection, even when selection might appear as the best explanation. In fact, in many cases, the patterns we observe emerge precisely because selection is not strong enough to overcome the effect of stochastic processes of mutation and genetic drift. Genomic signatures left by weak selection can mimic the footprints of adaptive evolution in several ways – from accumulation of intraspecific variation and accelerated divergence between species to strong biases in usage of alternative codons and amino acids. A detailed investigation of these sources of molecular information can, however, disentangle patterns emerging from adaptive and non-adaptive processes. Here, we use the social amoeba *Dictyostelium discoideum* as a model system to investigate fundamental evolutionary questions, with a special focus on disentangling the contribution of adaptive from non-adaptive forces shaping molecular variation. Chapter 1 provides a brief overview of the main points to be discussed throughout this work. In chapter 2, we integrate evolutionary theory with large-scale expression and genomic data from natural populations to understand evolutionary processes shaping genes associated with social behaviour. In chapter 3, we investigate implications of a strongly AT-biased genome for usage of alternative codons. In chapter 4 we address the often overlooked impact of overall processes shaping genome and cell economics on amino acid content and evolution of proteins. Finally, chapter 5 provides a short general discussion of the main findings of this work.

List of figures

Figure 1.1 The neutral theories of molecular evolution. 4
Figure 2.1 Polymorphism in social genes 29
Figure 2.2 The Direction of Selection (*DoS*) statistics for social genes 35
Figure 2.3 Evolutionary rates at social genes.....37
Figure 2.4 The impact of Red King processes on polymorphism and divergence39
Figure 3.1 Relative codon frequencies and GC content. 74
Figure 3.2 Nucleotide substitution matrix. 76
Figure 3.3 Contribution of neutral and adaptive processes to observed synonymous codon frequencies.....79
Figure 3.4 Patterns of inferred selection on codon usage bias. 81
Figure 3.5 Relationship between codon preference and expression optimization parameters. 82
Figure 3.6 Selection on overall GC content in coding regions. 84
Figure 4.1 Visualisation of the influence of base composition and metabolic cost on the use of amino acids. 100
Figure 4.2 Evolutionary signatures of genes with different levels of relative importance of base composition, cost and the interaction factor on overall amino acid usage. 102
Figure 4.3 Expression and strong selection shapes individuality of protein amino acid content. 104

List of tables

Table 1.1 Evolutionary signatures of genes underlying social interactions under different social dynamics 14
Table 2.1 Social genes in the social amoeba *D. discoideum* 28
Table 2.2 Tajima's *D* for social genes 30
Table 2.3 Enrichment analysis of the number of social genes carrying deleterious mutations 33
Table 4.1 Linear models explaining amino acid use across the genome 99

List of supplementary figures

Figure S2.1 Identification and characterization of sociality genes 54
Figure S2.2 Sliding widow analysis of differential expression.....55
Figure S2.3 Differential expression of *tgr* genes through development 56
Figure S3.1 Overall GC and transcript stability. 93
Figure S4.1 Fit of the model $M_{B+C+(B\times C)}$ of amino acid usage to proteins. 107
Figure S4.2 Correlation between biosynthesis costs and expression levels.....107

List of supplementary tables

Table S2.1 GO enrichment analysis for sociality genes.....57
Table S2.2 GO enrichment analysis for chimerism genes 58
Table S2.3 GO enrichment analysis for antagonism genes.....60
Table S2.4 GO enrichment analysis for cheater genes.....61
Table S2.5 Average number of SNPs (SNP/site) for social genes.....62
Table S2.6 Complementary neutrality tests for social genes 63
Table S2.7 Enrichment analysis of social genes evolving under balancing selection as defined by different cutoffs of Tajima's *D* 64
Table S2.8 Intraspecific variation in sociality genes excluding 13 genes evolving under balancing selection.....65
Table S2.9 Enrichment analysis of social genes showing strong signatures of selection.....66
Table S2.10 Evolutionary statistics for prespore and prestalk genes 67
Table S2.11 Enrichment analysis of the number of prespore and prestalk genes carrying at least one mutation that introduces a stop codon or results in a partial deletion (presence/absence variation) 68

Abbreviations

IGE	Indirect Genetic Effects
MKT	McDonald-Kreitman Test
DoS	Direction of Selection
RQ	Red Queen
ERQ	Escalatory Red Queen
FRQ	Fluctuating Red Queen
RK	Red King
CUB	Codon Usage Bias
tAI	tRNA Adaptation Index
StAI	Synonymous tRNA Adaptation Index

1 Introduction

1.1 A brief historical overview on the development of molecular evolution theories

Theoretical work developed by the Modern Synthesis laid the foundations of population genetics. Despite the inability to directly assay molecular variation, population genetics models from 'classical' and 'balancing' schools made different predictions on the amount of genetic variation present in nature. The former posited that most mutations are deleterious and rapidly removed, therefore polymorphism is expected to be low and transient; the latter posited that variation in nature would be high and caused by overdominant or frequency-dependent selection (reviewed in Nei 2013). Both schools agreed, however, that the main force driving evolution was natural selection (Page and Holmes 1998).

During the 1950-60s, the first insights into molecular variation emerged from studies of protein polymorphism (allozymes), which found levels of molecular variation to be very high, supporting predictions of the balancing hypothesis (reviewed in Page and Holmes 1998). However, these findings also posed a problem to this theory: if natural selection is the main force shaping genetic variation, there must be a great 'selective death' to remove all unfit individuals with inferior combinations of alleles ('cost of natural selection' (Haldane 1957)), which could, indeed, drive populations to extinction. These findings led Kimura (1968, 1985) to develop an alternative explanation. Because genetic variation is so high, a large fraction of it must be selectively neutral, and its levels reflect the net balance between mutation creating and genetic drift extinguishing neutral variation (Kimura and Ohta 1971). This means that polymorphism is transient: it is a 'momentary picture' of mutations captured in a certain point of space and time going through their journey to fixation or extinction by genetic drift. Consequently, evolutionary

rates reflect levels of intraspecific variation, and are solely determined by the rate of emergence of neutral mutations (Kimura and Ohta 1971).

Kimura's neutral theory (or 'simple neutral theory'[1]) (Kimura and Ohta 1971) revolutionized the field of population genetics because it was the first clear statement of a single mechanism for protein variation both within and between species, bringing together elements that were previously only weakly connected (Gillespie 1994). The simple neutral theory also provided the theoretical background for explaining the existence of a molecular clock (Kimura 1969), first anticipated by studies of protein polymorphism (Zuckerkandl and Pauling 1962). The reasoning is that, because rates of molecular evolution should reflect the rate of neutral mutations, which were considered to be constant over evolutionary time and across different lineages, the elapsed time since divergence of two lineages could be estimated in 'calendar time' (years) from the amount of genetic differences accumulated between them. However, evidences from experimental work challenged this view. First, estimates of evolutionary rates from DNA hybridization techniques (Laird et al. 1969) revealed a discrepancy between rates of nucleotide and amino acid substitutions, with only the former following the generation-time effect predicted by the simple neutral theory. This finding suggests that classes of nonsynonymous and synonymous substitutions do not respond to evolutionary forces of the same type and magnitude (Ohta and Gillespie 1996). Second, simple neutral theory predicts that the amount of intraspecific variation (polymorphism) scales with effective population size (N_e)[2] (Kimura 1968), but there is no apparent evidence of this. In fact, although population sizes differ by orders of magnitude across species, levels of genetic variation vary in a narrower range (Lewontin 1974). Finally, the assumption that most mutations are selectively neutral was difficult to

[1] Hereon, we adopt the nomenclature from (Ohta 1992) and refer to Kimura's Neutral Theory as 'simple neutral theory' as to distinguish from the 'nearly neutral theory'. Although they can be often referred to under the single term 'neutral theory' (Gillespie 1994), there are important distinctions between them that are focus of discussion throughout this chapter, thus the need to keep them as separate theories.

[2] Effective population size (N_e) is a central concept in population genetics, and it is often different from the census population size (N). N_e is the translation of N in a real population into the size of an idealized population showing the same rate of diversity loss as the real population under study (Wright 1931; Husemann et al. 2016). For simplicity, it can be interpreted as the number of individuals that actually contribute to the gene pool.

reconcile with growing evidence that amino acid substitutions result in microadaptations in different proteins and organisms (e.g. haemoglobin adaptations to different life styles and environments; this and other examples are compiled and discussed in (Gillespie 1994)).

Recognizing that Kimura's model was potentially oversimplified by classifying mutations as 'deleterious, neutral and advantageous' (Ohta 2012) (Figure 1.1A), Ohta extended simple neutral theory to account for the contribution of 'borderline' mutations (Ohta 1973, 1992): "natural selection cannot be so simple as to be 'all or nothing'" (Ohta 1992). Although developed as an extension to Kimura's neutral theory, some properties of the 'nearly neutral theory'[3] are in sharp contrast to the former. Namely, most mutations are considered to be slightly deleterious, instead of strongly deleterious or selectively neutral (Figure 1.1A). The destiny of these borderline mutations is categorically different from strictly neutral ones – it depends on the product of the effective population size (N_e) and selection strength (s), behaving as neutral when $N_e s$ is close to 0 (Figure 1.1B). Consequently, evolutionary rates are expected to decrease with effective population size, since selection is more efficient at removing slightly deleterious mutations (Ohta 1992; Ohta and Gillespie 1996; Akashi et al. 2012), precisely the opposite predicted by simple neutral theory (Kimura 1968).

[3] Nearly neutral theory and 'weak selection theory' (Akashi et al. 2012) are used interchangeably throughout this work.

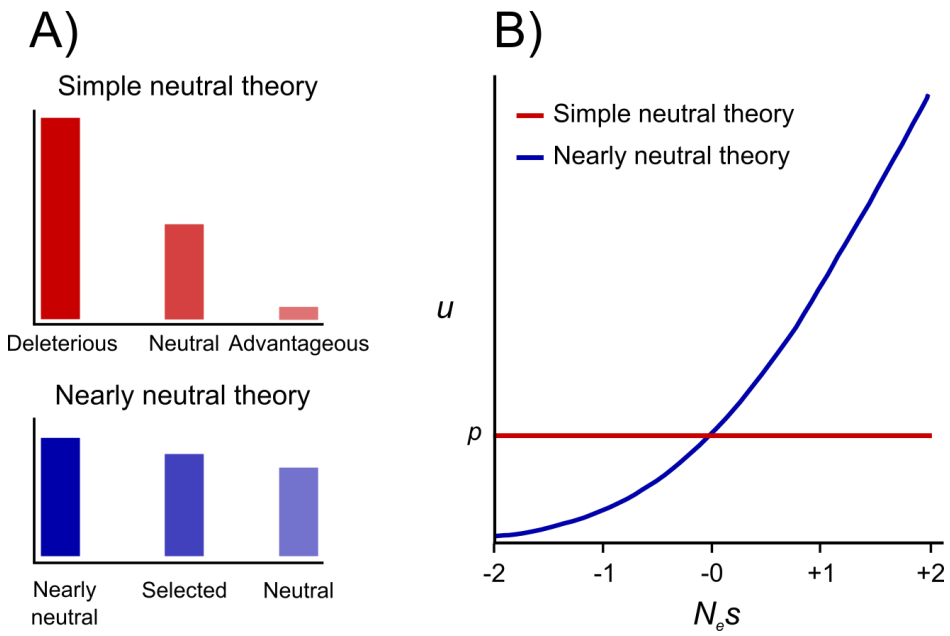

Figure 1.1 The neutral theories of molecular evolution.

A) Schematic plots showing the relative proportion of different classes of mutations under the two neutral theories. Deleterious mutants are definitely deleterious, and neutral mutants are strictly neutral. Most selected mutants are deleterious (selected against), but the group also includes advantageous alleles (selected for). Nearly neutral mutants comprise an intermediate class between neutral and selected mutants. **B)** The probability of fixation of a mutation (u) under the nearly neutral theory is a function of the product of the population size (N_e) and selection strength (s) ($N_e s$), whereas under the simple neutral theory it is the same across different values of $N_e s$ (p, the initial frequency). Figures in this panel were adapted (**A**) and redrawn (**B**) from (Ohta 1992).

Development of the nearly neutral theory had profound effects on evolutionary biology. One reason is that it provided a theoretical framework that integrates two competing evolutionary forces. Selection and genetic drift are no longer two categorical processes, but interplay on shaping genetic variation, where their relative contributions are ruled by the product of population size and strength of selection. Furthermore, whereas Kimura considered protein evolution to be governed by random processes independent of generation time, living conditions, and even morphological evolution (Kimura 1969), the nearly neutral theory provided the grounds to connect the stochastic view of sequence evolution to aspects of organismal biology. Ohta recognized that physiological conditions might influence weak selection, since constraints experienced by a protein can vary from

a biological system to another (Ohta 1992). Her theory also provided explanations for inconsistencies from molecular variation patterns to the neutral theory. For example, the apparent discrepancy between the relative constant rates of protein evolution and the variance on polymorphism levels at the DNA sequence (Laird et al. 1969), and the lower number of fixed differences in species with larger effective population sizes (Aquadro et al. 1988).

1.2 Weak selection and synonymous variation

In protein coding sequences, the effect of differences in selective constraints are often considered only at nonsynonymous sites, with synonymous sites generally considered to evolve neutrally (Kimura 1968; King and Jukes 1969). However, increasing evidence suggest that mutations at these sites are not as 'silent' as previously assumed. Although coding for the same amino acid, synonymous codons may differ in the availability of isoaccepting tRNAs carrying their particular anticodon (Ikemura 1981), which can, in turn, affect efficiency/accuracy of translation (Kurland 1992; Gingold and Pilpel 2011) or minimize protein misfolding (Drummond and Wilke 2009; Drummond et al. 2005). These signatures are magnified in highly and broadly expressed genes (Akashi and Eyre-Walker 1998) since they experience stronger selective constraints to optimize transcriptional/translational processes. Consistent with predictions from nearly neutral theory, synonymous codon optimization is broadly found in organisms with large effective population sizes (bacteria, yeast, bacteriophage and flies) (reviewed in Akashi and Eyre-Walker 1998), whereas biases in codon usage patterns are more influenced by base composition in organisms with smaller population sizes, such as vertebrates (Ikemura 1985).

But simple departures from equal usage of synonymous codons are not necessarily a sign of 'preference'. Patterns of codon usage bias across species can be strongly predicted by GC content from intergenic regions and local base composition (Chen et al. 2004). Similarly, in mammals, where GC content varies

widely across chromosome regions due to the presence of isochores (>>300 kb stretches of DNA with homogeneous base composition), synonymous codon usage is largely influenced by GC from surrounding regions (Bernardi et al. 1985; Bernardi 2000). Because GC content is often assumed to be determined by mutational pressures (Sueoka 1988, but see Rocha and Feil 2010), these findings suggest that strong patterns of codon usage bias can emerge due solely to background processes shaping the genome.

Building on nearly neutral theory, the question of how usage of alternative codons evolves seems to require an investigation not of which force generally *predominates*, but how different forces *interplay* in shaping variation at synonymous sites.

1.3 'Extended evolutionary null hypotheses'

Evolution under near neutrality is essentially dictated by the product $N_e s$, which turns this model into a fundamental null hypothesis for studying evolution in different lineages. In social systems characterized by division of labour, where reproduction can be bottlenecked to one or a few mating pairs, the effective population size is expected to be markedly reduced, since only reproducing individuals will pass their alleles to the offspring. In this regard, a comparative study revealed faster evolutionary rates associated with reduced efficiency of selection in social spiders of the genus *Stegodyphus* in comparison to sub-social conspecifics (Settepani et al. 2016). Sociality in spiders (including social species of the genus in that study) is associated with inbreeding, strong female biased sex ratios and reproductive skews (Lubin and Bilde 2007) – all factors that reduce effective population size (Wright 1931, 1932; Settepani et al. 2016).

Because expectations under the neutral and nearly neutral theories depend on effective population size to assess the role of drift in the evolutionary process, implications of these theories are often studied in the context of different groups (species or populations) or to understand demographic dynamics (e.g. population

expansions, bottlenecks, founder effects, etc.). But as revealed by a growing body of theoretical work, there are other ramifications of weak selection theory, accounting for differences in selective constraints across chromosomes (or chromosomal regions) and genes within the same genome.

Sex-biased or sex-exclusive inheritance of genome regions, as it is the case for animals with chromosomal sex determination, results in reduced effective population size in these regions compared to autosomal ones (Sayres 2018). In XY genetic systems, there are four copies of autosomes in a mating pair, but only three X and one Y chromosome. Thus, all else being equal, the reduced effective population size of sex chromosomes would result in faster evolutionary rates compared to autosomes (predicted to be even faster at the Y chromosome), only due to fixation of slightly deleterious mutations by genetic drift (Johnson and Lachance 2012). But all else is often not equal as far as sex chromosomes are concerned, and other factors are known to influence their evolution. When new (semi-) recessive mutations emerge on autosomes, they are often combined with the ancestral allele in a heterozygous genotype (because they are rare), so their effects are masked. However, when these types of mutations emerge on the X chromosome they are exposed to selection in the heterogametic sex (males) (Vicoso and Charlesworth 2006; Charlesworth, Coyne, and Barton 1987). As a result, selection on the X chromosome is predicted to be more effective than in autosomes (and evidence support this hypothesis (Mank et al. 2007; Lu and Wu 2005)), which is another reason to expect faster evolutionary rates on this chromosome, but now due to positive selection. Although this second scenario is not an evolutionary null hypothesis in a strict sense (neutral *versus* adaptive processes), it highlights the importance of considering factors that impact the efficiency of selection when investigating signatures of molecular evolution, particularly when other more complex scenarios are also plausible (such as sexual selection and sexual conflict).

Factors that decrease the phenotype-genotype association decrease, in general, the strength of selection, potentially resulting in elevated levels of segregating variation (Linksvayer and Wade 2009). For example, genes with indirect genetic effects (IGE), where fitness effects of a gene are not expressed in the focal individual carrying the gene, but in the phenotype of a different

conspecific individual (Wolf et al. 1998), can have a weaker genotype-phenotype relationship and therefore may experience weaker selection. For this reason, the expected consequences of this weaker association for the nucleotide sequence has been modelled in different groups of genes with IGE, such as those with maternal (Demuth and Wade 2007) and social (Linksvayer and Wade 2009) effects.

A potentially widespread source of IGEs arise from maternal effect genes, where genes carried by a mother have a causal influence in the offspring phenotype, irrespective of its genotype (Wolf and Wade 2009). Although post-natal influences such as the provision of food and choice of nesting sites in mammals and birds are the stereotypes of maternal effects, these effects can also arise from provision of nutrients and molecules that are pre-loaded in unfertilized eggs (Wolf and Wade 2009). For example, nutrients and mRNA molecules that are fundamental to early embryo development in flies are synthesized in the mother's nurse cells and transported to the oocyte during oogenesis (Schupbach and Wieschaus 1986; Spradling 1993). In *Drosophila*, this maternal provision includes transcription and transport of the major regulator of development of the anterior region: the gene *bicoid* (*bcd*) (Berleth et al. 1988). This maternal gene is only present in a derived group of flies (Cyclorrhapha) that includes *Drosophila*, and originated from an event of gene duplication of a *Hox3* gene (Stauber et al. 1999). In basal flies, *Hox3* has a maternal effect, just as *bcd*, but is also expressed later by the embryo genome. After the duplication event in the basis of Cyclorrhapha, each paralog assumed one of the functions previously performed by *Hox3* in lower Diptera (Stauber et al. 2002): *bcd* assumed the maternal role (indirect effect), whereas its paralog, *zerknült* (*zen*), assumed the zygotic role (direct effect). Taking advantage of this system, Demuth and Wade (2007) modelled the expected consequences of the indirect effect of maternally provided genes, such as *bcd*, in comparison to zygotic genes with direct effects, such as *zen*, and found that they are expected to evolve faster due to relaxed constraints. Elevated intraspecific variation and faster evolutionary rates of *bcd* in comparison to *zen* without departures from neutrality provide strong evidence to this hypothesis (Barker et al. 2005; Demuth and Wade 2007).

Genes with social effects form a particular class of genes with IGE, since they are both the targets and the agents of selection (Moore et al. 1997). Despite

their potentiality for a rapid adaptive evolution by reinforcement of some behavioural traits (e.g. aggression[4]), selection on genes involved with altruism experience selection only by kin selection (Hamilton 1964a, 1964b). This can be exemplified by insect societies where the workers provide care for the larvae, food and defence for the colony against intruders, but the role of reproduction is played out only by the queen. In such instances, genes from workers are not directly assessed by selection – because their fitness effects are expressed by individuals other than themselves –, but only indirectly, as a function of relatedness between these two classes of individuals. Consequently, genes underlying social traits are expected to harbour more intraspecific variation left behind by selection, and also to evolve faster due to fixation by genetic drift, in comparison to genes with direct effects (Linksvayer and Wade 2009).

This work on IGEs has been extended in a broader framework that has shown that selection can be weakened in a wider range of systems whenever the effect of a gene is conditional to a fraction of generations or to a subset of individuals within the same generation (Van Dyken and Wade 2010). In these cases, selection is expected to be weakened by a factor of $1/\phi$, where ϕ is the frequency of trait expression. This is even more critical if both cases occur at the same time in a system – i.e. if a trait is conditional to another condition. For example, assuming that, in the facultatively sexual species *Caenorhabditis elegans*, males appear once in every five generations and represent only 5% of the population, selection on a gene with male-limited expression is expected to be 1/100 of that experienced by a constitutively expressed gene (Van Dyken and Wade 2010). Pea aphids show a similar pattern, where females can reproduce asexually and males appear around once in each 10 or 20 generations. Following predictions from the theoretical model, investigation in this system has found signatures of weaker constraints in male-biased genes (Brisson and Nuzhdin 2008; Purandare et al. 2014).

These models provide adjusted evolutionary null hypotheses to account for peculiarities in modes of inheritance, genetic effects and frequency of expression

[4] Individuals carrying genes for aggressive behaviour turn the social environment more aggressive, which in turn may increase the selective pressure for more aggressiveness, and so forth.

by different groups of genes. Moreover, they highlight how, by weakening the strength of selection, these factors may leave signatures at the nucleotide sequence that resemble those left by different forms of selection (e.g. high polymorphism and rapid divergence by balancing and positive selection, respectively). This is particularly important when investigating evolutionary processes underlying the evolution of complex traits – such as many of those involved in social interactions –, since selective narratives are often suggested, but not contrasted against proper evolutionary null hypotheses (Nei 2005; Hughes 2008; Nei et al. 2010; Van Dyken and Wade 2012).

1.4 Sociogenomics: addressing social evolution at the molecular level

The genomics era inaugurated a new chapter in evolutionary biology, providing reliable large scale data to investigate long standing questions in population genetics, such as the determinants of standing polymorphism (Leffler et al. 2012; Ellegren and Galtier 2016) and the relationship between adaptive evolution and effective population size (Galtier 2016). Similarly, it has now provided the opportunity to understand the molecular mechanisms involved with expression of complex traits, such as social behaviour, as well as to assess the signatures of evolutionary processes recorded at the nucleotide sequences of genes underlying these traits. As a result, the social evolution literature has seen the emergence of a new field: Sociogenomics (Robinson 1999, 2002; Robinson et al. 2005).

Sociogenomic studies have the potential to integrate mechanistic and evolutionary analyses, in order to understand the molecular basis and evolution of social behaviours (Robinson 1999). For example, recent sociogenomic works have applied population genetics theory and molecular evolution tools to assess the evolutionary signatures of genes potentially involved with social behaviour, and concluded that they show signatures of conflict-driven evolution (Ostrowski et al.

2015; Noh et al. 2018). Although this framework can be very powerful to identify signatures of selection, it should be used with caution for several reasons. First, comparative studies have shown that evolutionary transitions – including a change from solitary to social life – are often addressed by changes in expression and network rearrangement of pre-existing genes (Kapheim et al. 2015; Glöckner et al. 2016). These findings suggest that the degree of pleiotropy in genes underlying social interactions is potentially high, so variation in these genes is likely shaped by factors other than their social role. Second, evolutionary tests may reveal patterns consistent with multiple evolutionary scenarios, thus requiring contrasts between multiple hypotheses, preferentially by performing tests that rely on different assumptions and uses different sources of data. For example, McDonald-Kreitman test (MKT) (McDonald and Kreitman 1991), which compares the proportion of segregating and fixed nonsynonymous and synonymous variation, may reveal a signature of balancing selection from an excess of nonsynonymous polymorphism, but it is very sensitive to slightly deleterious variation segregating under weak selection (Parsch et al. 2009). Similarly, Tajima's *D*, which compares the proportion of variation segregating at low and intermediate frequencies, can identify departures from neutrality by selection, but it is also largely influenced by demographic changes (Tajima 1989). Finally, but potentially more importantly, these complex evolutionary scenarios must be contrasted against appropriate evolutionary null hypotheses, considering the complexity and particularities of the system and genes under study.

Theoretical and experimental work on different social systems – from bacteria to ants – have revealed the importance of considering theoretical predictions from appropriate models as an evolutionary null hypothesis (Van Dyken and Wade 2012; Warner et al. 2017). In bacteria, quorum sensing genes are a well characterized group of genes involved with social traits (e.g. biofilm formation and bioluminescence), by producing signals that can be used by focal and surrounding cells (Miller and Bassler 2001). As a cooperative system characterized by joint production of a 'shared good', populations are constantly threatened by the emergence of cheaters (i.e. individuals that do not pay their fair cost but take advantage of the shared good). To understand the evolution of cheaters in this

system, Van Dyken and Wade (2012) first modelled theoretical expectations under the evolutionary null scenario that cheaters are transient because they emerge by recurrent mutation and are removed by purifying selection; and then, measured the intra- and interspecific levels of genetic variation on quorum sensing genes. Using this approach, they found high levels of variation both within and between species, following predictions of the null model, without the need to invoke the adaptive alternative hypothesis that cheaters are maintained by frequency-dependent balancing selection. Similarly, worker-biased genes, that are predicted to evolve under weaker constraints because they have indirect genetic effects (Linksvayer and Wade 2009), indeed show signatures of relaxed selection in ants (Warner et al. 2017). In fact, non-adaptive evolution seems to have even shaped genes on the onset of the transition from solitary to social life (Kapheim et al. 2015).

These contrasting results show how little we know about the evolution of genes underlying social traits so far, but also the diversity of scenarios that can be expected to shape these genes.

1.5 Identifying dynamics of social evolution from molecular evolutionary signatures

Cooperative social interactions typically require individuals to pay some cost, but the system is constantly threatened by invasion of cheaters that do not pay their fair share while reaping the benefits. Similar to interspecific conflicts (Brockhurst et al. 2014), such antagonistic interactions and symmetry in the strength of selection between interacting parties makes Red Queen (RQ) dynamics a likely process in the evolution of social interactions. One possibility is that there is an evolutionary cycle between cheating and resistance to cheating, which is analogous to the 'Escalatory Red Queen' (ERQ) process (Brockhurst et al. 2014), proceeding as a series of selective sweeps. Alternatively, the success conferred by a genetic variant may not be generalized across social contexts, but rather, may depend on the properties of the opponent or the specific context in which

competition occurs. Such non-transitivity can generate 'Fluctuating Red Queen dynamics' (FRQ) (Brockhurst et al. 2014), where frequency dependent selection maintains genetic variation and is manifested phenotypically as 'alternative strategies'. However, while considerations of social evolution often focus on the dynamic processes like the ERQ and FRQ, optimality approaches, such as game theory, predict a brake on the RQ processes: the appearance of a single unbeatable strategy (the 'evolutionarily stable strategy' – ESS) (Maynard-Smith and Price 1973). After such equilibrium arises, a period of evolutionary stasis is established, where variants that result in the emergence of new strategies are expected to be subject to purifying selection.

Factors other than conflict could also cause social genes to manifest different signatures of selection than other classes of genes. For example, social genes could actually be more dispensable if socially incompetent individuals suffered only a moderate loss of fitness. Such a scenario could arise because, in many systems, social genes might only be expected to experience natural selection or social selection some fraction of the time. Indeed, in organisms that are facultatively social, genes expressed only in social interactions may evolve under weak selection, since periods between social cycles should dilute the influence of selection arising from social interactions (Linksvayer and Wade 2009; Van Dyken and Wade 2010). These sorts of scenarios should lead to signatures of relaxed selection or potentially diminish any signatures from conflict (Linksvayer and Wade 2009; Van Dyken and Wade 2010; Linksvayer and Wade 2016). As a result, conditionally expressed social genes might be expected to harbour more variation or diverge faster than other genes simply because they experience weaker selection and hence are more subject to random drift (Linksvayer and Wade 2009; Van Dyken and Wade 2010; Linksvayer and Wade 2016).

Molecular population genetics tools can be applied to genes underlying social traits to distinguish competing hypotheses. Here, we compile a list of evolutionary tests that can be applied to disentangle the four social evolutionary dynamics discussed above (Table 1.1).

Table 1.1 Evolutionary signatures of genes underlying social interactions under different social dynamics

	ERQ	**FRQ**	**ESS**	**RK**
Main form of selection	Positive selection	Balancing selection	Purifying selection	Relaxed (diluted) selection
Polymorphism (SNP/site and π/site)	Low (new strategies are quickly selected and fixed by recurrent selective sweeps)	High and functional (nonsynonymous variation resulting in alternative strategies are maintained)	Low (variation is constantly removed by purifying selection)	High and include deleterious variation (variation accumulate as a result of poor/infrequent selection)
Range of excess of variation in the site frequency spectrum (e.g. Tajima's *D* test[a])	Lower and upper tails (segregating sites are either very close to fixation – when linked to a selected site –, or found at low frequencies – newly arising mutations); negative *D*	Intermediate (segregating variation is maintained at intermediate frequencies, inflating overall heterozygosity); positive *D*	Low (mutations are removed by selection before reaching higher frequencies); negative *D*	Distribution is closer to the neutral expectation (allele frequency decay with number of segregating sites); values of *D* into the range of neutrality, closer to 0.
Nonsynonymous polymorphism (P_n) relative to divergence (D_n) (MKT[b], Direction of Selection statistics[c])	$P_n < D_n$ (mutations resulting in new strategies are quickly fixed by selection, so they do not contribute much to polymorphism); positive *DoS*.	$P_n > D_n$ (variation is favoured when rare, but opposed when common, accumulating polymorphism that never gets fixed); negative *DoS*.	Although both P_n and D_n are usually low, the latter is always lower (depending on the strength of selection, mutations can segregate at lower frequencies, but is removed before getting fixed); negative *DoS*.	Both P_n and D_n are elevated, but the latter is usually lower (slightly deleterious mutations accumulate as polymorphism, and a fraction of it can be eventually fixed by drift; this fraction increases with dilution of selection); negative *DoS*, approaching 0 the more selection is weakened.

Rates of nonsynonymous substitutions (K_a[d])	Fast (mutations resulting in new strategies are quickly fixed by selection); $K_a/K_s > 1$.	Potentially slower (alternative alleles are maintained, but not fixed); $K_a/K_s < 1$.	Slow (new strategies are removed and do not spread in the population); $K_a/K_s << 1$.	Fast (mutations are fixed by drift more often); $K_a/K_s \sim 1$, if synonymous sites evolve neutrally.
Synonymous codon optimization (if there is evidence of it in the species under study)	No specific pattern	No specific pattern	No specific pattern	Higher segregation of non-optimal codons, increasing both polymorphism and divergence variation at synonymous sites.

[a] Tajima 1989

[b] McDonald and Kreitman 1991

[c] Stoletzki and Eyre-Walker 2011

[d] Nei and Gojobori 1986

1.6 *Dictyostelium discoideum* as a model system to understand evolutionary processes at the molecular level

Social behaviour in insects, birds and mammals has long called the attention not only of biologists, but also of other curious observers. This is presumably because such behaviours are very apparent to us: workers taking care of the brood and the hive, birds helping each other to wipe out parasites from the top of their heads, monkeys confabulating to ambush their prey. However, sociality goes beyond these classic examples, and even beyond organisms without anything resembling a brain, being described, for example, in a variety of microorganisms, such as viruses (Turner and Chao 1999), bacteria (Muñoz-Dorado and Arias 1995; Kirkup and Riley 2004) and amoebae (Strassmann et al. 2000). In fact, social microorganisms have emerged as important biological models in sociogenomic studies since their much simpler systems can help us identify genes underlying social interactions, as well as observe emergent dynamics of social evolution (Foster 2010).

The social amoeba *D. discoideum* lives as single-celled individuals feeding on bacteria in the soil, undergoing asexual vegetative growth like most microbes. However, when food is depleted, individuals aggregate in groups of about ~10^5 cells and go through a developmental cycle that ends with culmination of a fruiting body (Chisholm and Firtel 2004). Because of its peculiar life cycle, this amoeba has been widely used as an experimental model for investigations of cell signalling, morphogenesis and multicellular development (Kessin 2001; Chisholm and Firtel 2004; Eichinger et al. 2005; Rosengarten et al. 2015; Parikh et al. 2010), and to understand the transition from a uni- to multicellular lifestyle (Glöckner et al. 2016). More recently, interest on this system has also reached sociobiology (Strassmann, et al. 2000; Shaulsky and Kessin 2007; Li and Purugganan 2011). This is because aggregates may contain cells of different genotypes, resulting in a chimeric fruiting body, setting the stage for cheating: cheaters can exploit others strains and allocate more cells to the spore head than the fair proportion, without

contributing to form the sterile stalk (Strassmann et al. 2000). Besides the biological system itself, availability of a reference genome (Eichinger et al. 2005) and large scale expression data (Nasser et al. 2013; Parikh et al. 2010; Rosengarten et al. 2015), as well as a rich platform for genomic and experimental research (Fey et al. 2013), explain the emergence of this amoeba as a model species to dissect social evolution in a molecular level.

Investigations in this system have provided evidence for fundamental predictions of social evolution theory. For example, kin selection theory predicts that individuals cooperate as a function of relatedness, which can be addressed by population viscosity (leading to a high local concentration of closely related individuals) and/or development of a 'greenbeard', by which individuals carrying the same gene can be recognized by each other (Hamilton 1964a, 1964b). Relatedness is high among co-occurring natural strains, with estimates ranging from 0.52 in soil samples (Fortunato et al. 2003) to 0.86 in fruiting bodies (Gilbert et al. 2007). Evidence suggest that this can be due to growing as large clonal patches (Gilbert et al. 2009), but it is unlikely to be the only reason because not only population structure is considerably low in this system (Flowers et al. 2010), but to be a successful strategy in microbes, a strong population structure would require also selection to act simultaneously in multiple biological levels (Travisano and Velicer 2004). Instead, individuals developed mechanisms for genetic kin discrimination (Ostrowski et al. 2008), which is carried out by the pair of adhesion proteins *tgrB1* and *tgrC1*[5] (Benabentos et al. 2009) acting in a greenbeard-like manner (Gruenheit et al. 2017). Besides these two loci, several genes were identified as implicated in cooperation, either because disruption of these genes results in a cheating behaviour (Santorelli et al. 2008), or are differentially expressed during chimeric development (Li et al. 2014), or yet because they impose trade-offs by pleiotropic effects, stabilizing cooperation (Foster et al. 2004).

As a powerful social microbe model, recent works have also attempted to assess the evolutionary signatures of social dynamics in this system, and suggested a role for Red Queen processes (Ostrowski et al. 2015; Noh et al. 2018). However,

[5] Previously named *lagB1* and *lagC1*, respectively.

these studies used limited set of strains and tests, a few small groups of genes and, more importantly, lack a full consideration of appropriate evolutionary null hypothesis. As discussed on the previous sections, these factors can undermine our understanding of the real processes shaping genes underlying social behaviour, so studies incorporating these nuances are needed.

1.7 Aims and approaches

This study has the main aim of identifying the contribution of adaptive and stochastic processes on genome evolution in the social amoeba *Dictyostelium discoideum*. We start our work by characterizing evolutionary signatures of genes underlying social traits, by using large scale genome and transcriptome data, and contrasting competing evolutionary hypotheses. This is followed by an investigation of the processes shaping synonymous codon usage in this organism, under the null hypothesis that patterns can emerge from overall processes shaping the strongly AT-biased genome. Finally, we investigate the influence of background processes shaping genome and cell economics on amino acid content and protein evolution in this system, an effect that is often overlooked.

2 Molecular evolution of social genes is consistent with signatures of the Red King process in a microbe

Janaina Lima de Oliveira[1a], Atahualpa Castillo Morales[1a], Balint Stewart[2], Nicole Gruenheit[2], Jennifer Engelmoer[3], Suzanne Battom Brown[3], Reinaldo A. de Brito[4], Laurence D. Hurst[1], Araxi O. Urrutia[1], Christopher R. L. Thompson[2*], Jason B. Wolf[1*]

1. Milner Centre for Evolution and Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK

2. Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London, WC1E 6BT, UK

3. Faculty of Life Sciences, Michael Smith Building, University of Manchester, Oxford Rd, Manchester, M13 9PT, UK

4. Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, Brazil.

[a]These authors contributed equally to this work

[*]Correspondence to: jason@evolutionarygenetics.org & christopher.thompson@ucl.ac.uk

Keywords: Red Queen; conflict; cooperation; nearly neutral theory

2.1 Abstract

Social interactions are typically characterised by conflict and competition. This antagonism can play a critical role in evolutionary processes, such as promoting diversity through maintenance of alternative strategies or driving accelerated evolution through arms-race like escalation. However, despite our sophisticated understanding of how conflict shapes social traits, we still have limited knowledge of how it impacts molecular evolution across the underlying 'social genes'. To address this problem, we analysed the genome wide impact of social interactions in a microbe. Using genome sequences from 67 *Dictyostelium discoideum* strains, we find that social genes often exhibit enhanced polymorphism and accelerated evolution. However, these patterns are not consistent with the expectation of conflict driven processes, but instead reflect relatively relaxed purifying selection. This pattern reflects the fact that social interactions are conditional, and therefore selection on genes expressed in social interactions is diluted by generations of inactivity. This results in the 'Red King' process, wherein dilution of selection by inactivity enhances the role of drift, resulting in increased polymorphism and accelerated evolution.

2.2 Introduction

The social environment can have profound effects on fitness and, consequently, constitutes an important source of selection (Moore et al. 1997). It is generally thought that the social environment provides its most significant force of selection when interactions are characterised by conflict and competition. This is because antagonism can potentially generate a persistent, constantly changing source of selection, where social traits evolve in response to selection, and in turn change the nature of selection itself acting upon the genes underlying social traits (i.e., 'social genes') (Rice and Holland 1997). To date, however, research has largely focused on understanding how conflict driven selection affects the evolution of social traits, with the implied assumption that the underlying social genes would show similar patterns and processes. Thus, despite our sophisticated understanding of social trait evolution, we still have a limited understanding of how conflict ultimately impacts the social genes themselves (Robinson 1999; Robinson et al. 2005; Foster 2006). This is perhaps surprising given that the patterns of molecular evolution at social genes could help us better understand the key genes behind social traits, the nature of selection arising from social interactions, and the relative importance of different conflict driven processes in shaping social evolution.

The relentless selection resulting from social conflict is analogous to the Red Queen process, where competition in the ecological environment generates persistent counter-evolutionary change in interacting parties (Rice and Holland 1997; Brockhurst et al. 2014). The role of the Red Queen process in social evolution depends on the relationship between the selection imposed by social traits (where they are the agents of selection) and the corresponding selection experienced by social traits (where they are the targets of selection) (Moore et al. 1997). Hence, the consequences that these processes have on molecular evolution will, likewise, depend on the relationship between sequence variation at social genes and the properties of the social traits. One possibility is that selection favours constant evolutionary change in social traits, with reciprocal counter-evolution of competitive strategies akin to the 'Escalatory Red Queen' (Brockhurst et al. 2014). This process would presumably proceed as a series of selective sweeps of

advantageous mutations at the associated social genes, reducing levels of polymorphism and increasing the rate of evolutionary divergence. Alternatively, success in social interactions may depend on the specific properties of the opponent or context in which competition occurs, which could result in a scenario where different social traits, and hence genetic variants at associated social genes, are favoured in different social contexts. Such non-transitivity is akin to the 'Fluctuating Red Queen' process (Brockhurst et al. 2014), where frequency dependent selection maintains genetic variation that underlies alternative strategies, which would be manifested as a signature of balancing selection (Harris et al. 2008).

While conflict could potentially lead to the relentless Red Queen processes, with dramatic consequences for patterns of trait and molecular evolution, it can also potentially result in the opposite scenario, where evolutionary change is halted by the emergence an evolutionarily stable strategy (ESS) (Maynard-Smith and Price 1973). Populations at the ESS would experience optimizing selection to remain at the ESS, resulting in evolutionary stasis with purifying selection on associated social genes. Importantly, this purifying selection is expected to lead to low levels of polymorphism and divergence, which are at direct odds with the predictions of the Fluctuating and Escalatory Red Queen processes (Brockhurst et al. 2014). It is therefore possible to differentiate between contradictory predictions of conflict driven selection by evaluating signatures of selection on social genes, thus providing important insights into the nature and consequences of selection arising from social interactions.

Investigations into the form and consequences of selection generated by conflict must necessarily also consider the potentially confounding role of the random processes of drift and mutation (Linksvayer and Wade 2009; Van Dyken and Wade 2010, 2012). While this is true for all types of genes, it is particularly critical for social genes in organisms that are facultatively social or those that otherwise only rarely encounter conflict. Such 'conditionality' could dilute the impact of selection and enhance the role of drift (Linksvayer and Wade 2009; Van Dyken and Wade 2010). We refer to this scenario as the 'Red King' (RK) process. Unlike the Red Queen, who was constantly running, the Red King was mostly asleep in Lewis Carroll's "*Through the looking glass*", and hence the RK refers to

the impact of diluted selection owing to inactivity (or more generally, a lack of selection) under some conditions. Importantly, while both RK and Fluctuating Red Queen processes can potentially have similar consequences, such as elevated polymorphism in social genes, they typically differ in the specific signatures they predict. For example, the RK predicts an overall shift towards neutrality (manifested as an accumulation of slightly deleterious mutations) (Van Dyken and Wade 2010). In contrast, Fluctuating Red Queen is predicted to result in elevated functional variation (i.e., amino acid polymorphism) underlying adaptive alternatives (Brockhurst et al. 2014). Likewise, the RK process is expected to result in an elevated rate of fixation of slightly deleterious mutations by drift. Because most slightly deleterious mutations are nonsynonymous, greater fixation of nonsynonymous variation may resemble the positive selection favouring new strategies predicted for the Escalatory Red Queen. However, diluted selection under the RK process should allow for elevated levels of segregating deleterious polymorphism (Van Dyken and Wade 2010), while the selective sweeps of the Escalatory Red Queen would lead to lower levels of polymorphism (Brockhurst et al. 2014). Thus, in order to differentiate the impacts of different processes on molecular evolution, to ultimately understand how social interactions impact the genome, we need to consider the joint impacts of selection and drift on patterns of divergence and polymorphism at social genes.

Dictyostelium discoideum provides a powerful model system for studying the evolutionary consequences of social conflict (Strassmann et al. 2000) and for evaluating its impact on molecular evolution at social genes. *D. discoideum* live as single celled individuals in the soil, but aggregate together in response to starvation to form a multicellular slug that eventually forms a fruiting body that aids spore dispersal (Chisholm and Firtel 2004). Construction of a functioning fruiting body requires cooperation among cells, with some cells sacrificed to form the stalk, whilst others form viable spores. When multiple genotypes co-aggregate, this differentiation into stalk and spores is expected to be generate conflict over representation in spores (Strassmann et al. 2000). Previous analyses of social traits in this system have demonstrated enormous phenotypic diversity in traits associated with the social stage, including variation in relative representation in the sporehead,

spore size, and spore numbers (Buttery et al. 2009; Buttery et al. 2010; Wolf et al. 2015). This degree of phenotypic diversity suggests that evolutionary processes promote variation at social genes. However, a detailed analysis of social strategies suggests that there is potentially a single ESS, with facultative cooperation and cheating based on relatedness (Madgwick et al. 2018). Previous attempts to characterise patterns of molecular evolution at genes in *D. discoideum* did not reveal clear differential signatures of selection at social genes or were unable to distinguish between alternative hypotheses (Ostrowski et al. 2015; Noh et al. 2018). Therefore, to understand how social interactions have shaped gene sequence evolution, we have implemented an integrative approach using large-scale gene expression, functional genomics, and genome sequence data from 67 natural strains. By applying this approach to multiple sets of social genes identified through different complementary methods, we have been able to overcome past challenges to develop a clear picture of the evolutionary processes shaping signatures of selection at social genes. Our analyses provide strong support for a unified perspective, with all evidence consistent with the conclusion that social genes experience a similar overall pattern of selection as other classes of genes. However, because the expression of some social genes is restricted to the social stage, the patterns of molecular evolution manifest a signature of diluted selection owing to the Red King process.

2.3 Results

2.3.1 Identification of social genes

To understand broad-scale processes shaping molecular evolution at social genes we have used four different, but complementary, approaches to identify sets of social genes. For ease, we have named these sets ‘sociality’, ‘chimerism’, ‘antagonism’ and ‘cheater’ genes. For comparison, we have also identified appropriate sets of control genes.

2.3.1.1 Sociality genes

Sociality genes are defined as those with expression restricted to the social stage (which corresponds to the period of aggregation and multicellular development). Because sociality genes are only expressed in social stages, their evolutionary signatures should reflect the overall selective impact of social interactions. To identify sociality genes, we used large-scale transcriptome data from vegetative growth on bacteria or in liquid culture (Rosengarten et al. 2015; Nasser et al. 2013; Parikh et al. 2010) and high resolution transcriptome data from multiple stages of the social cycle, from starvation to the formation of mature fruiting bodies (Rosengarten et al. 2015). We calculated the Index of Social Expression (ISE) (Sucgang et al. 2011) for each gene by comparing the expression at 'social' stages (hour 1 to hour 24) to the expression of both social and single celled vegetative stages (hour 0). As expected, we find a clear discontinuity in the distribution of ISE values (Figure S2.1A). Importantly, 1650 genes exhibited a high bias in expression to social stages (ISE > 0.9; i.e. more than 90% of its expression concentrated in social stages), which we consider to be the set of sociality genes. Signatures of selection in sociality genes were compared against all genes with some level of expression in the full transcriptome dataset (i.e., all genes with some measured level of expression at any timepoint in development or in the vegetative stages).

Sociality genes were found to be expressed at remarkably low levels at the vegetative stage (median = -0.64 $\log_{10}$TPM, Figure S2.1B), demonstrating that they are effectively conditional to social development, and not simply up-regulated at this stage. Although not expressed in every generation, sociality genes are generally required at very high levels when expressed in the social cycle (Figure S2.1C). These genes are also overrepresented for GO categories related to development, such as culmination and sporulation (Table S2.1). Interestingly, this set is also enriched for genes without biological process annotation, which may reflect their lack of conservation and orthology with characterized genes, potentially reflecting rapid evolution.

2.3.1.2 Chimerism genes

Chimerism genes are defined as those up-regulated in chimeric aggregations in comparison to clonal aggregations. This is based on the logic that chimeric development will be characterised by conflict, and hence these genes will show the signatures of conflict driven evolution (Noh et al. 2018). To identify chimerism genes, RNA was extracted from aggregations composed from pairwise mixes of three wild strains after fourteen hours (corresponding to the slug stage) of clonal and chimeric development. RNA-seq revealed 190 genes showing significant up-regulation during chimeric development. These chimerism genes are enriched in GO categories mostly related to functions that are associated with vegetative growth, such as metabolic and biosynthetic processes (Table S2.2). Because these chimerism genes were identified from expression in clonal versus chimeric development, their evolutionary patterns were compared against all genes expressed in these contexts (i.e., all genes showing some level of expression under either condition).

2.3.1.3 Antagonism genes

Antagonism genes are defined as those genes that are preferentially expressed in cells destined to become the stalk or spores. These genes are candidates for those being shaped by antagonistic selection driven by conflict because cell fate choice in *D. discoideum* determines which cells end up having zero direct fitness by providing the dead stalk and which get the direct benefit by producing spores (Parkinson et al. 2011; Foster et al. 2004; Chattwood et al. 2013). Antagonism genes were identified as those that show differential expression in the cell populations in slugs that lead to the formation of the stalk (‘prestalk’ genes) and the spores (‘prespore’ genes) (Parikh et al. 2010; Noh et al. 2018). A total of 1901 genes show significant differential expression in either of these regions (prespore = 903 and prestalk = 998) (Noh et al. 2018). Antagonism genes are enriched in GO categories related to cell membrane, extracellular region and cytoskeleton (Table S2.3), which are potentially important to cell communication, cell sorting or morphogenesis. Signatures of selection in these genes were compared to the background of all genes expressed in prespore and prestalk cells (Parikh et al. 2010).

2.3.1.4 Cheater genes

A set of 99 cheater genes have previously been identified experimentally because they result in a 'cheater' phenotype (i.e. producing more than their fair share of spores) when mutated and mixed with wild type cells (Santorelli et al. 2008). Validation of these mutants in a fine scale was performed by recapitulating 10 insertional events by homologous recombination in wild-type cells. Phenotypes of these 11 mutants were identical with those of the original mutants in all cases. Because cheater genes were not identified by their expression profile, their evolutionary signatures can be compared to the rest of the protein coding genes in the genome. GO term analysis revealed that cheater genes are overrepresented in only one category of biological process: social behaviour (Table S2.4). However, this categorization appears to be tautological because it reflects genome annotation based on the mutagenesis screen used to identify these genes.

2.3.1.5 Overlap of social gene sets

Interestingly, there is little overlap between the sociality, cheater and chimerism sets of social genes (Table 2.1). Chimerism genes are not a subset of the sociality genes (see Table 2.1), and their mean ISE value is not significantly different from the rest of the genome ($ISE_{Chimerism} = 0.51$, $ISE_{Background} = 0.54$; t-test: *FDR*-corrected $P = 0.084$). In fact, we find that chimerism genes are actually significantly enriched for genes with the peak of maximum expression during vegetative growth (expected: 79; observed: 104; Chi-square test: $P < 0.0003$) and are enriched in GO categories mostly related to functions that are associated with vegetative growth (Table S2.2). Moreover, there is no significant overlap between cheater and sociality genes (Table 2.1). Although eight of the 99 cheater genes are expressed at such low levels during vegetative and developmental stages (across all sequenced RNA pools) that we cannot characterize their expression profile, the mean ISE for the remaining 91 genes is 0.53, which is not significantly different than all other genes ($ISE_{Cheater} = 0.54$, $ISE_{Background} = 0.53$; t-test: *FDR*-corrected $P = 0.740$). We also do not find an overlap between cheater and chimerism genes (Table 2.1).

Table 2.1 Social genes in the social amoeba *D. discoideum*

Sociality genes are effectively expressed in the social cycle, as measured by an index of social expression (see Methods). Chimerism genes are those differentially expressed in chimeras compared to clonal development, specifically at the slug stage. Antagonism is a group formed by previously identified genes differentially expressed in prespore and prestalk cells (Noh et al. 2018; Parikh et al. 2010). Cheater genes were previously characterized from mutagenesis screenings to identify mutants with defective behaviour (Santorelli et al. 2008). Significance for the overlaps between each pair of gene categories was obtained by chi-square tests. Observed and expected (in parenthesis) values are shown above the diagonal, and significant overlaps after *FDR* correction ($P < 0.05$) are highlighted in bold. These corrected *P*-values are shown below the diagonal.

	Sociality	Chimerism	Antagonism	Cheater
Sociality	—	22 (26)	**507 (261)**	9 (13)
Chimerism	0.543	—	**44 (30)**	2 (1)
Antagonism	**$< 10^{-70}$**	**0.015**	—	
Cheater	0.543	0.773	0.858	—

In contrast to the lack of overlap between the sociality, cheater and chimerism sets, we find that there is a significant enrichment in the overlap between antagonism genes and both sociality and chimerism genes (Table 2.1). This enrichment presumably reflects the fact that the antagonism genes were identified based on differential expression in slugs, and hence should show temporal overlap in expression with the sociality and chimerism genes (both identified from expression in social stages). This idea is supported by the fact that the antagonism genes show a significantly higher ISE than their respective background genes ($ISE_{Antagonism} = 0.63$, $ISE_{Background} = 0.53$; t-test: *FDR*-corrected $P < 10^{-48}$). This difference appears in both the prespore ($ISE_{Prespore} = 0.66$, $ISE_{Background} = 0.53$; t-test: *FDR*-corrected $P < 10^{-44}$) and prestalk ($ISE_{Prestalk} = 0.59$, $ISE_{Background} = 0.54$; t-test: *FDR*-corrected $P < 10^{-8}$) subsets, with a significantly higher index in the prespore set compared to prestalk (t-test: *FDR*-corrected $P < 10^{-7}$). Overall, these

patterns highlight the importance of taking multiple approaches to identify social genes.

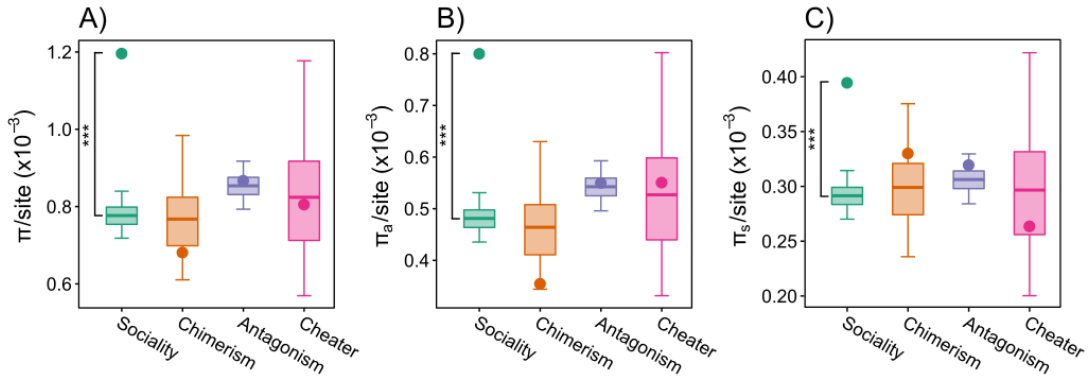

Figure 2.1 Polymorphism in social genes

Average estimates of nucleotide diversity per site (π/site) for CDS (**A**), nonsynonymous (**B**), and synonymous (**C**) sites for each group of genes (points) were compared to randomization distributions (boxplots). The middle line, bottom and top of the box show the expected mean, 25th and 75th percentiles respectively; whiskers present the 95% confidence interval of the distributions. Randomization distributions were generated for each group of social genes by generating a set of 10,000 random groups of genes of size *N* (where *N* corresponds to the number of genes in the particular group of social genes being tested). Randomization was done separately for each group of social genes by sampling from a set that contains that group of social genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance. Significance after *FDR* correction: $P < 0.05$ *; $P < 0.01$ **; $P < 0.001$ ***.

2.3.2 Only sociality genes harbour increased sequence diversity

To compare patterns of polymorphism in the four sets of social genes to their respective background gene pools, we generated genome sequence data from a set of 47 strains derived from the wild and combined these with sequence data from 20 published genomes (Gruenheit et al. 2017). Sequence information was obtained for 12,809 protein coding sequences. We find that only sociality genes differ in the levels of polymorphism from their background expectation, harbouring significantly more variation, whether estimated by average nucleotide diversity per site (π/site; Figure 2.1) or the number of SNPs per site (SNP/site; Table S2.5). Variation at sociality genes is greater across the entire CDS, including both

nonsynonymous and also synonymous sites. This pattern is consistent with the RK process, where signatures of relaxed selection are manifested at both synonymous and nonsynonymous sites, and inconsistent with the Fluctuating Red Queen process, where we expect signatures of balancing selection at nonsynonymous sites but not synonymous sites.

Table 2.2 Tajima's *D* for social genes

Expected values and the respective two-tailed *P*-values were obtained by a randomization process. For each group of social genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance after *FDR* correction for multiple tests.

Sites	Group	Expected	Observed	*P* (*FDR*)
CDS	Sociality	−0.629	−0.659	0.416
	Chimerism	−0.626	−0.654	0.914
	Antagonism	−0.628	−0.650	0.490
	Cheater	−0.628	−0.801	0.416
Nonsynonymous	Sociality	−0.614	−0.646	0.416
	Chimerism	−0.613	−0.633	0.914
	Antagonism	−0.614	−0.646	0.416
	Cheater	−0.613	−0.779	0.416
Synonymous	Sociality	−0.455	−0.451	0.914
	Chimerism	−0.456	−0.501	0.824
	Antagonism	−0.453	−0.450	0.924
	Cheater	−0.453	−0.591	0.416

2.3.3 Elevated variation in sociality genes reflects weak selection

To differentiate between alternative explanations for the pattern of elevated variation at sociality genes, we next examined the distribution of variation (i.e., the relative frequencies of polymorphisms) and the type of variation present (i.e., the relative frequency of deleterious variation, manifested as premature stop codons and partial gene deletions). For this we calculated average Tajima's *D* values (Tajima 1989) for each set of social genes, where negative values indicate an excess of low frequency variants (presumably reflecting erosion of variation by purifying selection or selective sweeps), and positive values an excess of intermediate frequency variants (reflecting maintenance of variation by balancing selection). The average *D* for sociality genes is negative for the whole coding sequence, as well as at nonsynonymous and synonymous sites when they are considered separately, but is not significantly different from that expected from the background genes (Table 2.2). This pattern is inconsistent with that expected under balancing selection and consistent with the expectation under either purifying selection or recent selective sweeps. This finding is supported by results from other neutrality tests, either using information from the site frequency spectrum (Fu & Li's *F** and *D**) or linkage disequilibrium statistics (Wall's *Q* and *B*) (Table S2.6).

To address the possibility that a subset of sociality genes experiences balancing selection, inflating the average polymorphism level for the group, we used two approaches. First, we tested whether sociality genes are enriched for genes evolving under balancing selection, using three different thresholds of *D* to define a signature of balancing selection ($D = 2$, $D = 1.5$ and $D = 1$). We find no evidence of such overrepresentation (Table S2.7). Regardless, there is a possibility that a small subset of sociality genes actually evolves under balancing selection and are responsible for the overall pattern of elevated nucleotide diversity within this group due to their hyper-variability. To evaluate this possibility, we identified 13 sociality genes evolving under balancing selection ($D > 2$), and removed them from the analysis of polymorphism. After this censoring, we still find that sociality genes exhibit significant higher levels of polymorphism (Table S2.8), supporting the

conclusion that the overall signature of selection on sociality genes is not a result of potential outliers under balancing selection.

To further differentiate between explanations for the elevated polymorphism at sociality genes, we focused on classes of segregating variation that are presumably deleterious. For this we examined the presence of two special types of mutations: SNPs that introduce a stop codon, and mutations that correspond to complete or partial gene deletion (which is characterised by presence-absence variation; PAV). We find that sociality genes are enriched for genes with both types of deleterious mutations (Table 2.3), which is consistent with relaxed purifying selection, thus providing further support for the RK. Interestingly, we also find that antagonism genes have a significant dearth of presence/absence variation, suggesting that they may be enriched for essential genes.

Table 2.3 Enrichment analysis of the number of social genes carrying deleterious mutations

We used a randomization procedure to test whether each of the five groups of genes contained an excess of genes carrying these types of deleterious mutations. For each group of genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. In each randomization we counted the number of genes that contained each type of deleterious mutation and used the distribution of the counts across randomizations to calculate the confidence intervals (2.5^{th} to 97.5^{th} percentiles) and *P*-values. Significant *P*-values after *FDR* correction for multiple tests are highlighted in bold (*FDR* < 0.05).

Class of mutations	Group	Observed	CI		*P* (*FDR*)
Stop codon gain	Sociality	79	46	72	**0.022**
	Chimerism	5	2	11	> 0.05
	Antagonism	11	4	15	> 0.05
	Cheater	9	1	8	> 0.05
Presence/Absence	Sociality	12	2	10	**0.042**
	Chimerism	0	0	2	> 0.05
	Antagonism	0	9	24	**0.002**
	Cheater	1	0	3	> 0.05

2.3.4 All classes of social genes primarily evolve under purifying selection

Analyses of the patterns of polymorphism provide only a partial picture of the nature of selection because different evolutionary processes can potentially result in similar levels of standing variation. For example, genes that show patterns of polymorphism that indicate purifying selection may also have recently experienced selective sweeps driven by the Escalatory Red Queen, since both processes erode variation. Therefore, we complemented our analyses of segregating polymorphism with two analyses that draw on patterns of evolutionary substitution to capture patterns of selection in deeper evolutionary time. Firstly, we compared levels of polymorphism to fixed differences in a highly divergent *D. discoideum* strain from Mexico (OT3A). Secondly, we characterized the rate of protein

sequence evolution by comparing the reference genome (Eichinger et al. 2005) to this divergent strain.

Using polymorphism data from the 67 natural *D. discoideum* strains and the divergent OT3A, we compared the number of segregating and fixed differences at each gene using the McDonald-Kreitman test (MKT) (McDonald and Kreitman 1991). The MKT identifies genes that have a significant excess of either nonsynonymous polymorphism (which could reflect either weak purifying or balancing selection) or nonsynonymous substitutions (reflecting positive selection for adaptive mutations). We found 47 genes that harbour a significant excess of nonsynonymous substitutions (D_n) and 94 showing an excess of nonsynonymous polymorphism (P_n). We next tested whether either of these classes of genes is enriched in any of the four groups of social genes in comparison to that expected for their comparable set of background genes. In sociality genes, we observe an underrepresentation of genes evolving under positive selection, suggesting a restricted role of adaptive evolution in this group (Table S2.9). For all other classes of genes, we find no significant overrepresentation of genes evolving under positive or balancing/purifying selection.

The MKT, which is based on a significant excess of either P_n or D_n, provides a conservative analysis and may not reveal subtle quantitative differences in evolutionary signatures. Therefore, we complemented the MKT with the Direction of Selection statistic: $DoS = D_n/(D_n + D_s) - P_n/(P_n + P_s)$ (Stoletzki and Eyre-Walker 2011). This approach provides a quantitative measure of the pattern of substitution relative to polymorphism, with zero indicating neutrality, positive values indicating adaptive evolution, and negative values indicating balancing selection or segregation of slightly deleterious variation. The average *DoS* value for sociality genes is negative, but not significantly different than expected compared to its background (Figure 2.2). However, the averages for both components of *DoS* – the proportion of substitutions ($D_n/(D_n + D_s)$) and polymorphisms ($P_n/(P_n + P_s)$) that are nonsynonymous – are higher in these genes, indicating both elevated variation and divergence. This pattern is inconsistent with the hypothesis that balancing selection maintains polymorphism. Instead, the overall pattern suggests weaker purifying selection on these genes, which leaves more deleterious mutational

variation segregating in the population and results in an increased probability of mutations eventually reaching fixation. For chimerism genes, the average *DoS* and the proportion of substitutions that are nonsynonymous are not significantly different from the background. However, they show a significant decrease in the proportion of polymorphisms that are nonsynonymous. For both the cheater and antagonism genes, the average *DoS* values, as well as the averages of both its constituents, are not significantly different from the background values. Taken together, quantitative *DoS* data suggests that all classes of social genes predominantly show patterns consistent with purifying selection, but vary in the relative intensity of selection.

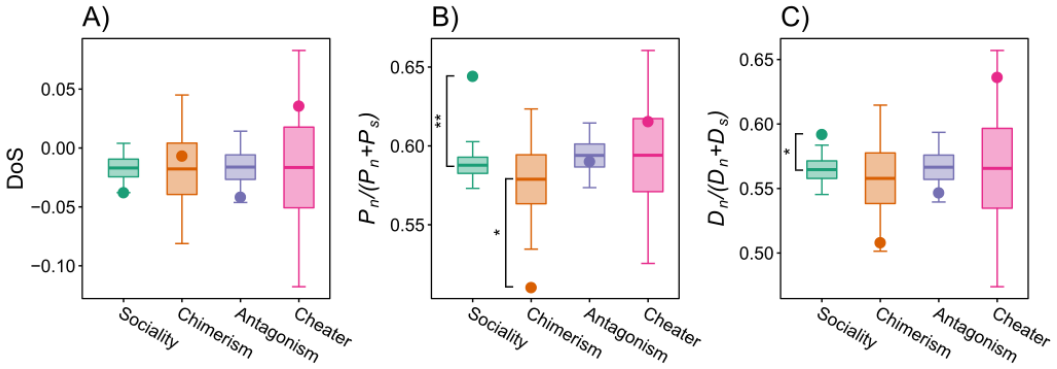

Figure 2.2 The Direction of Selection (*DoS*) statistics for social genes

Given is the *DoS* value (**A**), where $DoS = (D_n/(D_n+D_s))-(P_n/(P_n+P_s))$, and the value of each of its component parts: the proportion of polymorphisms $(P_n/(P_n+P_s))$ (**B**) and substitutions $(D_n/(D_n+D_s))$ (**C**) that are nonsynonymous. Average estimates for each group of genes (points) were compared to randomization distributions (boxplots). The middle line, bottom and top of the box show the expected mean, 25th and 75th percentiles respectively; whiskers present the 95% confidence interval of the distributions. Randomization distributions were generated for each group of social genes by generating a set of 10,000 random groups of genes of size *N* (where *N* corresponds to the number of genes in the particular group of social genes being tested). Randomization was done separately for each group of social genes by sampling from a set that contains that group of social genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance. Significance after *FDR* correction: $P < 0.05$ *; $P < 0.01$ **; $P < 0.001$ ***.

Rates of protein evolution were calculated from pairwise gene alignments using the reference genome (Eichinger et al. 2005) and OT3A. The number of

nonsynonymous substitutions per nonsynonymous site (K_a) was compared to the number of synonymous substitutions per synonymous site (K_s) to identify signatures of selection. The ratio K_a/K_s is expected to be ~1 if nonsynonymous sites are nearly neutral, > 1 if they are under positive selection, and < 1 if they are under purifying selection (Hurst 2002). We identified 11,901 protein coding orthologues between this pair of lineages, and estimated K_a and K_s. We next removed all orthologues where the ratio could not be calculated (for example, when synonymous sites are saturated). Using the remaining 5509 genes, we then tested whether the four classes of social genes differed from their respective backgrounds. For all sets of social genes, the average K_a/K_s is < 1, but patterns varied across the classes (Figure 2.3). Cheater and antagonism genes do not differ from their respective backgrounds for any parameter (K_a, K_s or K_a/K_s). For chimerism genes, however, both K_a and K_s are significantly lower than expected, while K_a/K_s is not different from expected (Figure 2.3). In contrast, sociality genes show increased substitution rates at both nonsynonymous and synonymous sites. However, because both classes of substitutions change in the same direction, the overall rate of evolution (K_a/K_s) does not differ from the background rate. Although we would expect the Escalatory Red Queen process to lead to accelerated rates of evolution at protein coding genes, these results are not consistent with the Escalatory Red Queen expectations because we would not expect the observed corresponding rates of evolution at both synonymous and nonsynonymous sites. Hence, these findings support the initial hypothesis that both synonymous and nonsynonymous sites in social genes evolve under purifying selection, but with the RK process reducing the strength of selection that results in an increased rate of sequence evolution. Thus, we find that all classes of social genes appear to evolve under a similar overall pattern of purifying selection, but with chimerism genes experiencing the strongest selection, cheater and antagonism genes an intermediate value, and sociality genes the weakest selection.

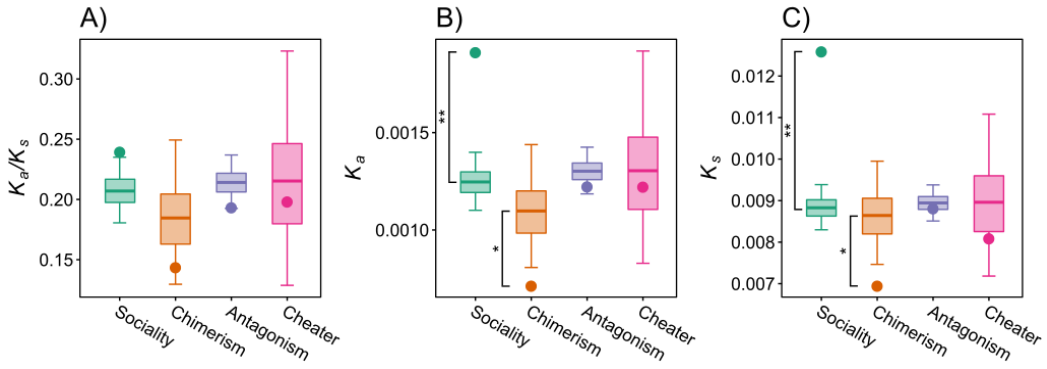

Figure 2.3 Evolutionary rates at social genes

Given is the evolutionary rates (**A**), and rates of substitution at nonsynonymous (K_a) **B**) and synonymous (K_s) sites (**C**). Substitutions represent changes compared to the sequence of a divergent strain, OT3A, from Mexico. Average estimates for each group of genes (points) were compared to randomization distributions (boxplots). The middle line, bottom and top of the box show the expected mean, 25th and 75th percentiles respectively; whiskers present the 95% confidence interval of the distributions. Randomization distributions were generated for each group of social genes by generating a set of 10,000 random groups of genes of size *N* (where *N* corresponds to the number of genes in the particular group of social genes being tested). Randomization was done separately for each group of social genes by sampling from a set that contains that group of social genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance. Significance after *FDR* correction: $P < 0.05$ *; $P < 0.01$ **; $P < 0.001$ ***.

2.3.5 The RK process as a unifying explanation for patterns of molecular evolution

The patterns of polymorphism and divergence are all consistent with the hypothesis that each class of social genes evolves primarily under purifying selection, with differences being explained by the degree of conditionality leading to the RK process. To test whether the RK process provides an overall explanation for patterns of molecular evolution, we examined the relationship between the overall degree of conditionality for a gene class (i.e., the proportion of sociality genes in the class) and either the levels of polymorphism or divergence. To improve the power and resolution of the analysis, we added five more classes of genes to increase sample size and expand the coverage of different degrees of conditionality: non-sociality (which show some level of expression in the transcriptome dataset

used to identify sociality genes, but which are not conditional to the social stage), non-chimerism (genes expressed in clonal and/or chimeric slugs, but which are not up-regulated in chimeric slugs), non-antagonism (genes expressed in prestalk and prespore cells and show no differential expression in these two cell types), and two classes of antagonism genes showing differing degrees of differential expression in prestalk versus prespore cells (representing expression biases of 0.8 and 0.9 corresponding to 480 and 105 genes, respectively). We do not include the non-cheater genes since that set essentially represents all protein coding genes. Using a weighted regression to account for the variation in the sizes of the gene classes (based on the $\sqrt{N}$ for genes included in the calculation of the relevant statistic), we find that the degree of conditionality accounts for the vast majority of the variation in patterns of polymorphism in terms of π/site (R^2 = 0.93, 0.90 and 0.85 for the full CDS, nonsynonymous and synonymous sites respectively, with $P < 0.001$ in all cases; Figure 2.4). Conditionality also accounts for the majority of the variation in the rate of synonymous (K_s, R^2 = 0.90, $P < 0.001$) and nonsynonymous (K_a, R^2 = 0.67, $P < 0.01$) divergence, but does not explain variation in the rate of nonsynonymous relative to synonymous divergence (K_a/K_s, R^2 = 0.24, P = 0.12). This pattern suggests that conditionality is facilitating the fixation by drift of slightly deleterious mutations, which appears as a constant proportional rate of both synonymous and nonsynonymous substitution. The resulting linear models that relate proportion of conditionally expressed genes to the various evolutionary parameters (Figure 2.4) provide a means of predicting the evolutionary signatures at a group of genes based solely on the proportion of member genes that are conditionally expressed. Hence, they account for variation in the importance of nearly neutral processes and, therefore, provide a null expectation for each set of genes, which corresponds to the RK prediction. As a result, for any category to be considered as having patterns that are inconsistent with this null hypothesis, they would have to differ significantly from the expectation from the regression. In the case of the classes of genes we have analysed, none differ significantly from the RK prediction, strongly suggesting that the variation in evolutionary signatures observed in other studies (Ostrowski et al. 2015; Noh et al. 2018) are artefacts introduced by not accounting for the RK as the appropriate null hypothesis.

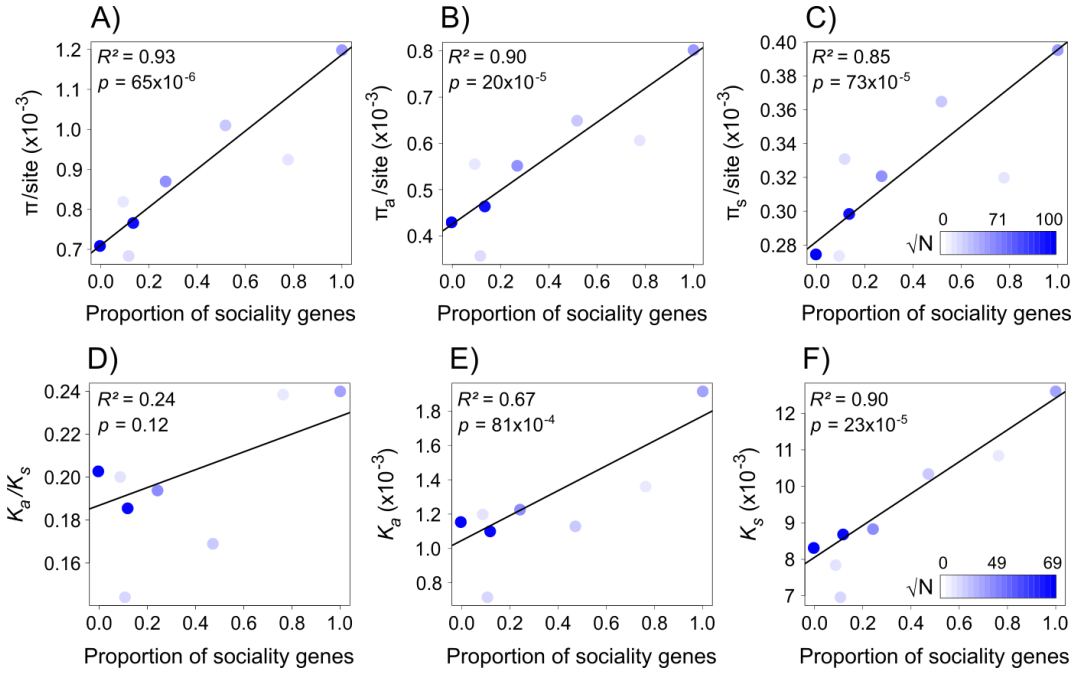

Figure 2.4 The impact of Red King processes on polymorphism and divergence

The top row of panels shows the relationship between nucleotide diversity in social genes at the CDS (**A**), nonsynonymous (**B**) and synonymous sites (**C**) as a function of the proportion on conditionally expressed (sociality) genes within that group. Similarly, the bottom row of panels shows the relationship between the rates of nonsynonymous (**E**), synonymous substitutions (**F**), and the rate of protein evolution (**D**) in a group of genes as a function of the proportion of sociality genes in each group.

2.4 Discussion

Social environments differ from other environments because they are created by interactions between individuals and hence are a property of the population. Therefore the social traits that form the social environment can be both the agents and the targets of selection (Moore et al. 1997). Theory has shown that this phenomenon can lead to very different evolutionary dynamics for social traits, and hence at their underlying genes, compared to other types of traits. From a genetic perspective, this reciprocal relationship between the source and targets of selection means that genes can generate selection on one another (including

themselves), leading to the potential for concerted coevolution (or counter-evolution) of genes with the sources of selection they experience (Moore et al. 1997; Rice and Holland 1997). For example, concerted evolution of traits and the social environment could lead to a runaway process that results in exaggeration of social traits, such as those involved in elaborate displays (Moore et al. 1997; West-Eberhard 1979). While studies at the phenotypic level have demonstrated broad support for theoretical expectations of how social traits evolve (Dugatkin 1998; Turner and Chao 1999; Foster 2010; Bijma et al. 2007), we still have a surprisingly limited empirical understanding of how evolution plays out at the underlying 'social genes' (Robinson 1999; Robinson, Grozinger, and Whitfield 2005; Foster 2006). Because the ultimate evolutionary impact of social interactions at the level of the genome should be recorded in the nucleotide sequence of underlying 'social genes', the evidence manifested in the signatures of selection on gene sequences can potentially reveal the fundamental properties and impact of the social interactions driving selection.

By implementing a broad set of analyses of molecular evolution across several complementary sets of social genes in the social amoeba *D. discoideum*, our study provides strong support for a unified evolutionary picture. Overall, social genes appear to experience a similar pattern of selection as other genes, which is primarily characterised by purifying selection. However, where we do find differences in their evolutionary patterns, these are consistent with a scenario in which conditional expression and the relative use and disuse of genes plays a critical role in shaping evolutionary patterns, which we term the Red King process. For example, chimerism genes show a significantly lower proportion of polymorphisms that are nonsynonymous (Figure 2.2B), even though levels of overall polymorphism are similar to control groups of genes (Figure 2.1B). This signature of relative stronger evolutionary constraints at these genes (in comparison to their comparable background genes) is also manifested in the lower rate of both nonsynonymous (Figure 2.3B) and synonymous (Figure 2.3C) divergence. The apparent constraint on chimerism genes likely reflects the fact that they are enriched for genes with their maximal expression during vegetative growth and have a small number of conditionally expressed genes. In fact, their average ISE is very close to 0.5

(indicating that they are, on average, expressed at similar levels in both the vegetative and developmental stages). In contrast, for sociality genes, which have expression that is essentially restricted to the social stage, we see a very clear picture of relatively weaker purifying selection (compared to their background genes), both in the level of polymorphism and divergence (Figures 2.1 to 2.3). Thus, while overall patterns primarily reflect purifying selection, the likely explanation for why sets of genes still differ in their manifested molecular signatures of selection is that they experience selection with varying frequencies (i.e., they have different degrees of conditionality). It is interesting to note that this conclusion bears some similarities to the finding that breadth of expression is a good predictor of the strength of purifying selection on mammalian genes (Duret and Mouchiroud 2000). However, there is a critical difference between how breadth of expression and the RK process impact molecular evolution. Breadth of expression is likely to reflect the extent of pleiotropy for a gene, with higher pleiotropy (associated with broader expression) generally leading to stronger selection owing to the accumulated effects of selection across tissues or traits. In contrast, when expression is restricted to a subset of generations, selection is diluted by conditionality, irrespective of the strength of selection during the generations where expression occurs. Hence, breadth of expression impacts molecular evolution because it is a determinant of the strength of selection, whereas the RK process impacts molecular evolution by shifting the overall distribution of selection coefficients towards zero.

These results highlight that, when testing evolutionary hypotheses, it is critical to consider the impact of the RK process and identify the correct null hypothesis for why genes might show different signatures of selection. Besides conditionality, expression level is another variable to be taken into account before interpreting signatures of molecular evolution, since highly expressed genes in general evolve slower (Drummond et al. 2005). However, the weaker selection experienced by sociality genes seems to result from conditionality rather than lower expression, since although restricted to a fraction of generations, expression of these genes is generally high when required (Figure S2.1C). Moreover, our analyses emphasize the importance of evaluating the full body of evidence when interpreting patterns of molecular evolution, since incorrect conclusions can be drawn from the

results of any single test. For example, we see elevated polymorphism at sociality genes, which could be misclassified as evidence for balancing selection. However, the fact that we see commensurate increases in both nonsynonymous and synonymous polymorphism, including different classes of deleterious mutations, and no significant difference from the background genes in the value of Tajima's *D* strongly supports the conclusion that the increased polymorphism reflects weaker purifying selection, not balancing selection.

A key challenge for studies aimed at understanding the molecular evolution of social genes is in first identifying representative social genes (Robinson et al. 2005; Robinson et al. 2008; Foster et al. 2007). This is exacerbated by the fact that different methods can all potentially introduce biases. To solve these problems, we have utilised tractability of microbial systems for the discovery of social genes. Moreover, we used four different approaches to identify largely independent groups of genes. Remarkably, despite the fact that the groups of social genes were independently identified as being associated with social interactions, no gene was identified by all four methods, and for the chimerism, sociality and cheater genes, there is no significant overlap between these classes (i.e., no significant enrichment, see Table 2.1). Antagonism genes show significant enrichment in their overlap with sociality and chimerism genes, presumably because they must be expressed during the slug stage (which is the same developmental timepoint at which chimerism genes were identified). However, despite the significant enrichment, these classes still only contain a relatively small proportion of shared genes (Table 2.1). Although the groups differ in some of their evolutionary signatures, they all fit cleanly into the unified predictions under the RK process. This conclusion is clearly captured in Figure 2.4, which demonstrates that different categories of genes all fall in the overall patterns of polymorphism (Figures 2.4A-C) and divergence (Figures 2.4D-F) predicted based on their degree of conditionality (i.e., the proportion of their genes that are conditionally expressed only in social stages). Most remarkably, the degree of conditionality explains the large majority of variation across gene classes in the levels of polymorphism and divergence, with no group appearing as a significant outlier in this overall pattern. Thus, despite the fact that previous studies of variation at social genes in this species have suggested that social genes

show signatures consistent with patterns driven by social interactions (Ostrowski et al. 2015; Noh et al. 2018), we see no evidence in support of this conclusion.

Many molecular evolution studies begin with the a priori assumption that synonymous substitutions are neutral. However, throughout our analyses we see strong evidence that variation at synonymous sites is under selection. Consequently, synonymous sites provide a critical additional body of evidence for evaluating the relative strength of selection, which clarifies the evolutionary picture. Importantly, synonymous sites are unlikely to be sources of functional (and potentially adaptive) variation, such as that underlying different social strategies (which presumably arises primarily from nonsynonymous differences in genes). Therefore, synonymous variation is unlikely to be maintained by balancing selection, and hence is most likely to reflect inefficient purifying selection driven by codon use or other processes in transcription and translation (such as splice control). Thus, the fact that we typically see differences among groups of genes in their evolutionary signatures (for both polymorphism and divergence) at both synonymous and nonsynonymous sites suggests the same phenomenon is affecting the strength of selection at all sites in the CDS. This result is most clear in Figure 2.4, where we see a similar dependence of the levels of polymorphism and divergence at both nonsynonymous (Figures 2.4B and 2.4E, respectively) and synonymous sites (Figures 2.4C and 2.4F, respectively) on the degree of conditionality. The fact that all of these values change in the same way strongly suggests that they all reflect the dilution of selection on conditionally expressed genes.

Although we have focused on the impact of direct selection shaping variation and divergence at social genes, other analyses of how social interactions impact molecular evolution have considered the impact of kin selection (Noh et al. 2018). In the case of *D. discoideum*, the impact of kin selection could potentially be seen in differences in selection acting on genes expressed in cells destined to become stalk (which potentially experience kin selection) or spores (which presumably experience direct selection). It has previously been suggested that this scenario results in the dilution of selection on prestalk genes owing to the indirect nature of kin selection relative to prespore genes (Noh et al. 2018), and thus consistent with kin selection. However, our analyses, which are based on a much

larger set of genome sequences, provide several lines of evidence that strongly counter the conclusion that kin selection has left a signature in the patterns of molecular evolution at these genes. Most importantly, we find no evidence that the two sets of genes differ in their patterns of sequence evolution (hence our combining them here), nor do they differ from the relevant background genes (Tables S2.5-7, S2.9-11). This is perhaps unsurprising since we find no evidence that genes expressed in either of these cell populations show exclusive expression in either of these conditions. Moreover, most of these genes are also expressed across different stages. Therefore, despite their significant expression bias in the prestalk and prespore regions, these genes are not expected to differ in the relative importance of direct and kin selection.

In summary, we find that the Red King process, wherein the relative use and disuse of social genes across generations modulates the relative strength of selection they experience, provides a unifying explanation for large-scale evolutionary patterns. This conclusion does not rule out a role for other evolutionary processes, like the RQ, at some genes, but the impact is likely restricted to a relatively small collection of genes or sites. In the context of social conflict, the overall pattern of purifying selection at genes associated with the social stage (regardless of how they were identified) is consistent with there being an overall optimum, as expected under an ESS, but that selection is diluted due to conditionality. Given that phenotypic studies have identified conspicuous differences between naturally occurring strains in all traits measured (Buttery et al. 2009; Buttery et al. 2010; Wolf et al. 2015), our results suggest that the observed variation potentially reflects the inefficiency of selection to remove variation, rather than selection maintaining a diversity of alternative strategies.

2.5 Methods

2.5.1 Genomic DNA sequencing

Genomic DNA was extracted and sequenced from 58 *D. discoideum* strains and one divergent Mexican *Dictyostelium* strain (OT3A, which is characterised as *D. discoideum*, but could represent a close congener), all obtained from the Dicty Stock Center (Fey et al. 2013). For DNA extraction, 10^9 cells were collected after growth on nutrient media that contained *Klebsiella aerogenes*. Cells were re-suspended in KK2 and washed at least three times by centrifugation at 2200 rpm for 2 minutes to remove remaining bacteria. Nuclei were extracted from the pellet containing amoeba, followed by genomic DNA extraction as described elsewhere (Gruenheit et al. 2017). gDNA quality was assessed by agarose gel electrophoresis and a NanoDrop spectrophotometer (Thermo scientific). gDNA was quantified using a Qubit® fluorometer (Thermo scientific) before genomic libraries were prepared using Illumina TruSeq kit. Paired-end sequencing for reads ranging from 75-100 bp were obtained on an Illumina HISeq sequencer. A second round of library sequencing was performed for strains NC105.1, DD185, K10, S109, QS102, NC85.2 and NC60.1 in order to increase the number of reads. To complement our de novo sequencing we also downloaded raw reads from NCBI Sequence Read Archive (SRP071575) of published genome sequence data from another 20 *D. discoideum* natural strains (Gruenheit et al. 2017) (Table S2.12).

2.5.2 Mapping and SNP calling

Reads where cleaned for adapters and quality trimmed using Trimmomatic (Bolger et al. 2014) allowing maximally 2 mismatches in seed alignments and extending and clipping if a score of 30 is reached. Leading and trailing bases with a quality less than 3 where removed, before scanning the reads with a 4-base sliding window and cutting if the average quality per base drops below 15. Reads with a length of less than 36 bases after this process where then dropped. In order to separate *D. discoideum* reads from those derived from possible contaminants,

trimmed reads where binned by simultaneously mapping them to the reference genome of *D. discoideum*, *Paraburkholderia xenovorans lb400*, *Burkholderia ubonensis*, *Paraburkholderia fungorum* and *Klebsiella pneumoniae*; and assigning them according to the best mapping score using BBSplit, part of the BBMap package (Bushnell 2016). Genomes from the aforementioned bacterial species where downloaded from Ensembl Bacteria database (Kersey et al. 2016). Reads binned with *D. discoideum* or not mapped in the previous step where pooled together and mapped to the *D. discoideum* reference genome using NextGenMap (Sedlazeck et al. 2013).

SNP calling was performed by comparison with the reference genome (Eichinger et al. 2005) using the Genome Analysis Toolkit GATK (McKenna et al. 2010), following Best Practices recommendations for standard hard filtering parameters (DePristo et al. 2011; Van der Auwera et al. 2013). Briefly, alignments where sorted and PCR duplicates marked using Picard tools (Wysoker et al. 2016). Base quality score recalibration (BQSR) was performed by calling SNPs in each strain, filtering out sites with a Quality lower than 30, depth of coverage lower than 2, quality by depth (QD) less than 2, Fisher strand bias (FS) over 60 or Mean Mapping Quality (MQ) less than 40. Remaining SNPs where then used to perform BQSR using GATK. Variants where then jointly called on the 79 strains using GATK HaplotypeCaller and GenotypeGVCFs functions. Resulting SNPs where filtered with a static threshold of QD < 2.0 || FS > 60.0 || MQ < 30.0. As to maximise the number of informative sites for posterior analysis, while reducing the amount of noise introduced by missing genotypes in strains with low genome coverage or high diversity, we removed any strain with a missing call rate higher than 0.3, any site called in less than 90% of the remaining strains (i.e. in less than 60 out of 67 strains), as well as any multiallelic site or indel. This results in a dataset of 279,807 SNPs across 67 strains.

2.5.3 Intraspecific evolutionary statistics

Parameters of genetic diversity (number of SNPs and the average nucleotide diversity, π) and Tajima's *D* where estimated for genes with an average mapping of

more than 50%, using the R package PopGenome (Pfeifer et al. 2014). The two diversity measures were estimated for coding regions, nonsynonymous and synonymous sites, and then scaled to the mapped CDS length to obtain per site measures. Characterization of SNPs that introduce premature stop codons was performed by using SnpEff (Cingolani et al. 2012). Genes with an average mapping ≤ 50% were considered to hold a presence/absence variation (PAV) and were analysed separately to assess if this type of structural genetic variation is more frequent among any group of social genes.

2.5.4 Interspecific divergence

SNPs were further characterized as nonsynonymous (n) or synonymous (s) and segregating (P) or fixed (D) differences by comparison to a Mexican *Dictyostelium* isolate OT3A. While this strain is annotated as *D. discoideum* in dictyBase (Fey et al. 2013), the low mapping rate of our sequencing reads and the high divergence of this strain with respect to all other isolates suggest that this strain belongs to a different species, or at the very least, is an outgroup to the strains used in this study. We used this information to perform the McDonald-Kreitman test using the R Package PopGenome (Pfeifer et al. 2014). These counts were also included in the calculation of the Direction of Selection (*DoS*) statistic (Stoletzki and Eyre-Walker 2011). In both cases, the analysis was conducted for each gene individually, not by pooling all SNPs from genes pertaining to the same group.

To calculate rates of protein evolution we compared the reference genome of *D. discoideum* (Eichinger et al. 2005) to OT3A. We first built the pseudo genome of OT3A by inserting SNPs for this strain (with comparison to the reference genome) into the reference genome, by using VCFtools software package (Danecek et al. 2011). CDSs for all genes from both genomes where extracted using gffread (Pertea 2017) and rates of synonymous (K_s) and nonsynonymous substitutions (K_a) where estimated using R package seqinR (Charif and Lobry 2007). The rate of protein evolution K_a/K_s was calculated for each CDS and averaged for alternative transcripts of the same gene.

2.5.5 Identification of social genes

2.5.5.1 Sociality genes

To identify genes biased to the social (developmental) cycle of *D. discoideum*, we used data from several published RNA-seq experiments sampled from vegetative growth (Rosengarten et al. 2015; Nasser et al. 2013; Parikh et al. 2010) and from the developmental transcriptome (Rosengarten et al. 2015; filter experiment). In total, we used data from seven vegetative conditions (15 replicates) and 18 developmental time points during the social stage sampled at every 1-2h (from hour 1 to hour 24, 2 replicates each) (Rosengarten et al. 2015). Reads were downloaded from NCBI Gene expression Omnibus (GEO: GSE61914), trimmed with skewer package (Jiang et al. 2014) and filtered for a minimum length of 20bp and a mean Phred Quality score of 20. Remaining reads where pseudo-aligned to transcripts of the *D. discoideum* reference genome (Eichinger et al. 2005) downloaded from Ensembl Protists database release 36 (Kersey et al. 2016) and further quantified using Kallisto (Bray et al. 2016). One hundred bootstrap samples were generated for each replicate to compute uncertainty estimates for the expression levels. Normalisation was performed using the TMM method (Robinson and Oshlack 2010) implemented in edgeR (Robinson et al. 2010) and scaled to coding sequence length, after discarding genes with less than two reads in less than two libraries.

Our analyses of differential expression across time points (Figure S2.2) agreed with previous findings that genes up-regulated one hour following starvation have GO categories consistent with a shift to multicellular social development (Rosengarten et al. 2015). We also find that the *tgr* genes, which are known to play an important role in social interactions (Benabentos et al. 2009; Gruenheit et al. 2017) are up-regulated at this stage (Figure S2.3). Consequently, we considered the social stage to begin at the first hour. Therefore, data from hour 1 to hour 24 are considered to be part of the 'social library', while data hour zero and from all vegetative conditions are considered to be part of the 'vegetative libraries'.

In order to define sociality genes, we averaged values for replicates and calculated an index of social expression (ISE), defined as the proportion of the total expression that appears in the social libraries (Sucgang et al. 2011):

$$ISE = \frac{\overline{Social\ libraries}}{\overline{Vegetative\ libraries} + \overline{Social\ libraries}}$$

Sociality genes were defined as those with an index higher than 0.9.

2.5.5.2 Chimerism genes

To identify genes showing differential expression under chimeric conditions we experimentally created clonal and all pair-wise chimeric aggregations using three strains originating from the same geographical location (NC34.2, NC57.1 and NC87.1). Cells of each strain were grown in association with *Klebsiella aerogenes*, before washing by centrifugation in KK2 buffer. Washed cells were then plated on non-nutrient L28 purified agar (agar) at a density of x cells/cm^2. For chimeric combinations we mixed equal numbers of cells from each strain. Aggregations were harvested after fourteen hours of development, when slugs had formed. This stage was chosen because previous work has demonstrated that the effects of chimeric development can be seen at this stage (Foster et al. 2002; Castillo et al. 2005; Jack et al. 2015; Gruenheit et al. 2017). Development of each clone and chimeric combination was carried out in duplicate. For each replicate, slugs from ten agar plates were pooled for RNA extraction using Trizol. RNA pools were sequenced on an Illumina TruSeq with 100 bp paired-end reads following standard protocols. This yielded between ~10^7 to 2 x 10^7 (mean ~1.5 x 10^7) reads per RNA pool.

Preprocessing and mapping was performed as described above for the identification of sociality genes. Briefly, reads were trimmed and filtered using the skewer package (Jiang et al. 2014) (min. length of 20bp, mean Quality score of 20). They were then pseudo-aligned to *D. discoideum* transcripts (Eichinger et al. 2005) obtained from Ensembl Protists database release 36 (Kersey et al. 2016) and quantified using Kallisto (Bray et al. 2016). One hundred bootstrap samples were generated for each replicate to compute uncertainty estimates for the expression levels. Genes with less than 5 reads in at least 47% of the libraries were discarded.

Estimates of expression were then summarized to gene level and Wald test for differential expression was performed for chimeras and clonal samples by using sleuth (Pimentel et al. 2017). Chimerism genes are then defined as those that are significantly up-regulated in chimeric slugs (*FDR* adjusted *P*-value < 0.05).

2.5.5.3 Antagonism genes

A list of 903 prespore and 998 prestalk genes was obtained from ref. (Noh et al. 2018), which derived from an RNA-seq experiment identifying genes differentially expressed in these two cell subtypes (Parikh et al. 2010). For evolutionary analyses, these genes were compared against all genes in the expression data provided in ref. (Parikh et al. 2010). One prespore and four prestalk genes in the prespore/prestalk list were not present in the original data, but were included in our analysis in both: the background and the specific groups of genes.

For the regression analysis of the impact of Red King processes on polymorphism, we included two extra sets of antagonism genes based on their expression bias between prestalk and prespore regions of the slug (corresponding to biases of ≥ 0.8 and ≥ 0.9 in either cell type). For this, we combined the data from ref. (Parikh et al. 2010) with an additional set of data, which was generated as follows. *D. discoideum* cells transformed with either ecmAO-RFP or pspA-RFP reporter genes (Parkinson et al. 2009) were developed to the slug stage. Slugs were collected in dissociation buffer (KK2, 10mM EDTA) and dissociated through a G21 needle. Cells were resuspended at 10^8 cells/ml and cell clumps removed by filtration. RFP expressing cells were purified using a BD FACSaria flow sorter. Total RNA was extracted using TRI Reagent, before rRNA depletion using Ribminus^{TM} Eukaryotic kit (Invitrogen). 200-500ng of rRNA depleted RNA was reverse transcribed, fragmented and size selected for 150-250 bp cDNA fragments. cDNA was amplified using strand specific primers and sequenced using a SOLiD 4 system. Expression biases were calculated separately for each dataset as the proportion of the total expression that appears in the prestalk libraries compared to prespore libraries, and vice-versa, as follows:

$$Prestalk\ bias = \frac{\overline{Prestalk\ libraries}}{\overline{Prestalk\ libraries} + \overline{Prespore\ libraries}}$$

$$Prespore\ bias = \frac{\overline{Prespore\ libraries}}{\overline{Prestalk\ libraries} + \overline{Prespore\ libraries}}$$

Genes were included in each set if the two datasets agreed on the degree of bias (e.g., if the bias calculated from both sets was ≥ 0.8 the gene would have been included in the 0.8 bias set).

2.5.5.4 Cheater genes

Previous work has identified mutations that result in a facultative reduction of cooperative behaviour when *D. discoideum* strains grown in chimeras with a different strain (Santorelli et al. 2008). These genes where identified by screening of insertional mutagenesis (REMI) libraries, and a subset of mutants was validated at a finer scale by recapitulating 11 insertional events (10 intra- and 1 intergenic mutations) by homologous recombination in wild-type cells. Phenotypes of these 11 mutants were identical with those of the original mutants in all cases. A fraction of these mutations, occurring in intergenic regions, were discarded. The remaining mutations affect a total of 99 genes, which we referred to as 'cheater' genes.

2.5.6 GO enrichment

GO terms for biological process, cellular component and molecular processes were obtained from Dictybase (Fey et al. 2013). Enrichment analyses of GO categories in sociality, chimerism, antagonism and cheater genes were performed in R and statistical significance was assessed after *FDR* adjustment of one tail *P*-values from Monte Carlo sampling (see below).

2.5.7 Randomization procedure for significance testing

All statistical analyses and data manipulation were performed in R version 3.3.0 and RStudio version 0.99.902, using built-in functions and the package ggplot2 (Wickham 2016) for plotting. Unless otherwise explicitly stated in the text, significance was assessed by randomization tests. For each evolutionary analysis, 10,000 samples (*s*) of the same size of the group of genes being tested (sociality, chimerism, antagonism or cheater genes) were taken. Each sample was averaged (continuous variables) or the number of genes showing a particular feature was computed (categorical variables) – both cases resulting in a distribution of 10,000 random samples. Expected values were provided as the mean of this random distribution. Significance of categorical variables were first assessed by comparing observed values to the 95% confidence interval (CI). When these values lie outside the CI, numeric two tail *P*-values were calculated as twice the number of times that the observed count for the particular group of genes did not exceed the one in the randomly generated subset divided by 10,000. Two tail *P*-values for continuous variables were obtained similarly, but using averages values instead of counts. For every statistical test, an *FDR* (Benjamini-Hochberg) correction for multiple tests was performed.

2.5.8 Data availability

All data generated or used in the current study are publicly available. The list of genetic variants used in all analyses are available from the EMBL-EBI European Variation Archive (EVA) (IDs pending and available on request). The transcriptome data used in the analysis of sociality genes were downloaded from NCBI Gene expression Omnibus (GEO: GSE61914). The list of prespore and prestalk genes used in the analysis of antagonism genes was obtained from Noh et al. (2018), which was combined with a list of all genes included in the original RNA-seq experiment from Parikh et al. (2010). The list of cheater genes is available from Santorelli et al. (2008). The RNA-seq (transcriptome) data sets from the comparison of clonal and chimeric slugs (used in the analysis of conflict genes) and from the comparison of prestalk and prespore regions (used to identify genes with biased expression in these regions for the linear model testing the effect of

proportion of sociality genes) are available from the NCBI Gene Expression Omnibus (both IDs pending and available upon request).

ACKNOWLEDGEMENTS: This work was funded by grants from the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/M01035X/1; BB/M007146/1) to J.B.W., L.D.H., and C.R.L.T., the Natural Environment Research Council (NE/H020322/1) to J.B.W. and C.R.L.T., the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) to R.A.B., a Wellcome Trust Investigator Award (WT095643AIA) to C.R.L.T, a BBSRC studentship to S.B., and a studentship from CNPq (234216/2014-0) to J.L.O. We thank: Mike Wade for critical discussions that motivated this work, Joan Strassmann and Dave Queller for discussions during the development of this work, and Carlos Congrains for help with the implementation of the computational components.

AUTHOR CONTRIBUTIONS: J.L.O. and A.C.M performed all of the statistical analyses of molecular evolution, A.C.M. performed all of the processing of genome sequence data and analyses of the gene expression data, J.L.O., A.C.M., L.D.H., A.U. R.A.B., and J.B.W. conceived the analyses, B.S. and N.G. contributed to the genome sequencing pipeline, C.R.L.T. designed the gene expression experiments, J.H. performed the gene expression experiment in clonal versus chimeric slugs (for chimerism gene identification), S.B. performed the gene expression experiment comparing prestalk and prespore regions, C.R.L.T. and J.B.W. conceived the project, J.L.O., A.C.M., L.D.H., A.O.U., C.R.L.T., and J.B.W. wrote the paper.

2.6 Supplementary material

Figures

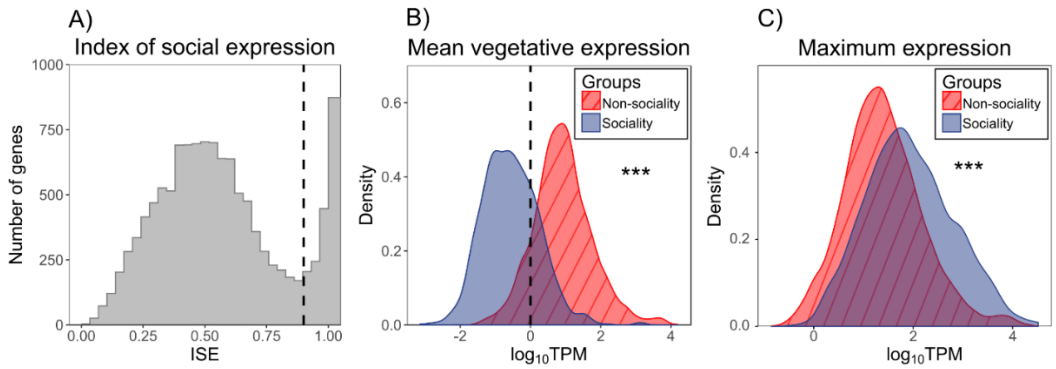

Figure S2.1 Identification and characterization of sociality genes

A) Distribution of values for the Index of Social Expression (ISE). The dashed line represents the cutoff of ISE = 0.9 used to define sociality (ISE > 0.9) and non-sociality (ISE ≤ 0.9) genes. **B)** Sociality genes have little or no expression during vegetative growth (Kolmogorov-Smirnov test: $P < 10^{-15}$), suggesting that they are conditional to the social stage. **C)** Although conditional to a fraction of generations, sociality genes are usually required at high levels when expressed (Kolmogorov-Smirnov test: $P < 10^{-15}$).

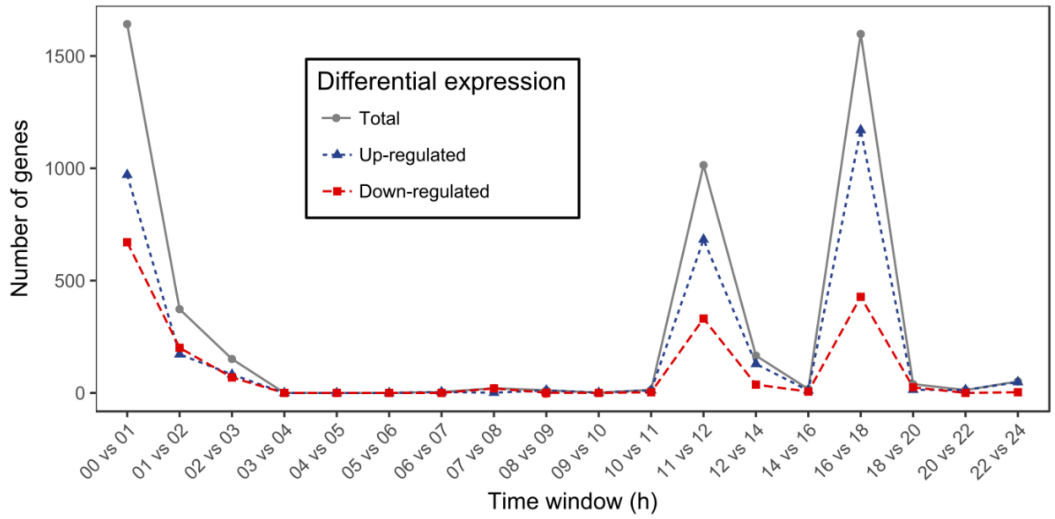

Figure S2.2 Sliding widow analysis of differential expression

By computing the number of differentially expressed genes between a given time point and the subsequent one (*t* versus *t+1*), an analysis of the developmental transcriptome reveals three major points of global changes in expression patterns. The first step marks the beginning of development (00-01h), suggesting that conditional expression of developmental genes is observed as early as within the first hour of starvation. The second and third peaks are related to switches from loose aggregates to multicellularity (11-12h) and beginning of culmination (16-18h), respectively (see also Rosengarten et al. (2015)).

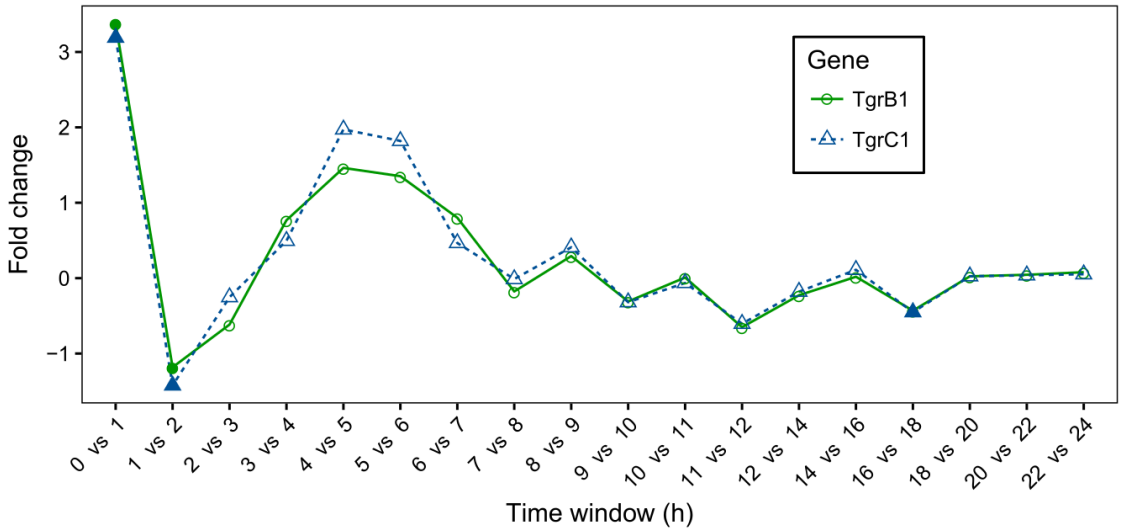

Figure S2.3 Differential expression of *tgr* genes through development

The pair of developmental genes *tgrB1* and *tgrC1* is up-regulated (filled symbols, positive fold change) on the onset of development, between the vegetative stage and the first hour of starvation. They are further down-regulated between hours 1 and 2, and again at the beginning of culmination (hours 16 and 18) (filled symbols, negative fold change). In other time points, transcripts of these genes are accumulated and increase levels, but are not differentially expressed (empty symbols).

Tables

Table S2.1 GO enrichment analysis for sociality genes

We used a randomization procedure to test whether this group of genes is enriched for GO terms of biological process, cellular component and molecular function. For each GO term, we generated a set of 10,000 random groups of size *N* (where *N* is the number of sociality genes) sampled from a set that contains sociality genes and its corresponding background set of genes. In each randomization we computed the number of genes associated to the GO term being tested and used the distribution of the counts across randomizations to calculate the one-tail *P*-values. Only terms overrepresented among sociality genes after *FDR* correction are shown.

GOID	GO Term	Obs	Exp	*P*	*FDR P*
	Biological Process				
GO:0030198	extracellular matrix organization	29	4.6	$<10^{-4}$	$<10^{-4}$
GO:0030435	sporulation resulting in formation of a cellular spore	24	10.26	$<10^{-4}$	$<10^{-4}$
GO:0031154	culmination involved in sorocarp development	28	12.71	$<10^{-4}$	$<10^{-4}$
GO:1902168	response to catechin	7	1.38	$<10^{-4}$	$<10^{-4}$
GO:0008150	biological_process	411	263.12	$<10^{-4}$	$<10^{-4}$
GO:000NABP	no biological process annotation	803	695.64	$<10^{-4}$	$<10^{-4}$
	Cellular Component				
GO:0005576	extracellular region	83	26.82	$<10^{-4}$	$<10^{-4}$
GO:0016021	integral component of membrane	369	308.21	$<10^{-4}$	$<10^{-4}$
GO:0031012	extracellular matrix	30	4.84	$<10^{-4}$	$<10^{-4}$
GO:0005575	cellular_component	438	318.09	$<10^{-4}$	$<10^{-4}$
GO:000NACC	no cellular component annotation	702	630.71	0.0002	0.0264
	Molecular Function				
GO:0004497	monooxygenase activity	27	8.62	$<10^{-4}$	$<10^{-4}$
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	25	9.18	$<10^{-4}$	$<10^{-4}$
GO:0005201	extracellular matrix structural constituent	27	4.41	$<10^{-4}$	$<10^{-4}$
GO:0005506	iron ion binding	23	9.73	$<10^{-4}$	$<10^{-4}$
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or	21	6.96	$<10^{-4}$	$<10^{-4}$

	reduction of molecular oxygen				
GO:0020037	heme binding	26	10.02	$<10^{-4}$	$<10^{-4}$
GO:0030246	carbohydrate binding	43	14.42	$<10^{-4}$	$<10^{-4}$
GO:0001646	cAMP receptor activity	6	1.11	0.0001	0.0159
GO:0030248	cellulose binding	28	14.05	0.0003	0.0433
GO:0003674	molecular_function	370	268.91	$<10^{-4}$	$<10^{-4}$
GO:000NAMF	no molecular function annotation	786	669.99	$<10^{-4}$	$<10^{-4}$

Table S2.2 GO enrichment analysis for chimerism genes

We used a randomization procedure to test whether this group of genes is enriched for GO terms of biological process, cellular component and molecular function. For each GO term, we generated a set of 10,000 random groups of size *N* (where *N* is the number of chimerism genes) sampled from a set that contains chimerism genes and its corresponding background set of genes. In each randomization we computed the number of genes associated to the GO term being tested and used the distribution of the counts across randomizations to calculate the one-tail *P*-values. Only terms overrepresented among chimerism genes after *FDR* correction are shown.

GOID	GO Term	Obs	Exp	*P*	*FDR P*
	Biological Process				
GO:0006096	glycolytic process	6	0.29	$<10^{-4}$	$<10^{-4}$
GO:0006099	tricarboxylic acid cycle	7	0.40	$<10^{-4}$	$<10^{-4}$
GO:0006108	malate metabolic process	4	0.10	$<10^{-4}$	$<10^{-4}$
GO:0006164	purine nucleotide biosynthetic process	5	0.18	$<10^{-4}$	$<10^{-4}$
GO:0006338	chromatin remodelling	4	0.28	$<10^{-4}$	$<10^{-4}$
GO:0006457	protein folding	12	1.21	$<10^{-4}$	$<10^{-4}$
GO:0006458	'de novo' protein folding	5	0.29	$<10^{-4}$	$<10^{-4}$
GO:0006471	protein ADP-ribosylation	3	0.09	$<10^{-4}$	$<10^{-4}$
GO:0006520	cellular amino acid metabolic process	5	0.18	$<10^{-4}$	$<10^{-4}$
GO:0006531	aspartate metabolic process	2	0.03	$<10^{-4}$	$<10^{-4}$
GO:0006532	aspartate biosynthetic process	3	0.05	$<10^{-4}$	$<10^{-4}$
GO:0006536	glutamate metabolic process	3	0.09	$<10^{-4}$	$<10^{-4}$
GO:0008152	metabolic process	24	5.67	$<10^{-4}$	$<10^{-4}$
GO:0009408	response to heat	4	0.15	$<10^{-4}$	$<10^{-4}$
GO:0031589	cell-substrate adhesion	5	0.46	$<10^{-4}$	$<10^{-4}$
GO:0046689	response to mercury ion	6	0.76	$<10^{-4}$	$<10^{-4}$
GO:0055114	oxidation-reduction process	22	6.70	$<10^{-4}$	$<10^{-4}$
GO:0061077	chaperone-mediated protein folding	5	0.44	$<10^{-4}$	$<10^{-4}$

GO:0000492	box C/D snoRNP assembly	2	0.03	0.0001	0.0092
GO:0006189	'de novo' IMP biosynthetic process	3	0.11	0.0001	0.0092
GO:0030435	sporulation resulting in formation of a cellular spore	7	1.18	0.0001	0.0092
GO:0070212	protein poly-ADP-ribosylation	2	0.03	0.0001	0.0092
GO:0008652	cellular amino acid biosynthetic process	4	0.32	0.0002	0.0168
GO:0019752	carboxylic acid metabolic process	3	0.17	0.0002	0.0168
GO:0006807	nitrogen compound metabolic process	3	0.18	0.0003	0.0216
GO:0010421	hydrogen peroxide-mediated programmed cell death	3	0.14	0.0003	0.0216
GO:0010918	positive regulation of mitochondrial membrane potential	3	0.14	0.0003	0.0216
GO:0019538	protein metabolic process	2	0.03	0.0003	0.0216
GO:0006094	Gluconeogenesis	3	0.16	0.0004	0.0269
GO:0009617	response to bacterium	6	0.99	0.0004	0.0269
GO:0000398	mRNA splicing, via spliceosome	6	1.04	0.0005	0.0315
GO:0046956	positive phototaxis	3	0.17	0.0005	0.0315
GO:0006538	glutamate catabolic process	2	0.05	0.0007	0.0415
GO:0046847	filopodium assembly	3	0.21	0.0007	0.0415
GO:0006734	NADH metabolic process	2	0.05	0.0008	0.0448
GO:0051103	DNA ligation involved in DNA repair	3	0.17	0.0008	0.0448
GO:0006273	lagging strand elongation	3	0.19	0.0009	0.0491
	Cellular Component				
GO:0005634	Nucleus	33	16.65	$<10^{-4}$	$<10^{-4}$
GO:0005681	spliceosomal complex	6	0.94	$<10^{-4}$	$<10^{-4}$
GO:0005737	Cytoplasm	61	19.66	$<10^{-4}$	$<10^{-4}$
GO:0005739	Mitochondrion	19	6.13	$<10^{-4}$	$<10^{-4}$
GO:0005759	mitochondrial matrix	7	1.05	$<10^{-4}$	$<10^{-4}$
GO:0005829	Cytosol	21	5.27	$<10^{-4}$	$<10^{-4}$
GO:0045335	phagocytic vesicle	34	4.87	$<10^{-4}$	$<10^{-4}$
GO:0097255	R2TP complex	2	0.03	$<10^{-4}$	$<10^{-4}$
GO:0005832	chaperonin-containing T-complex	5	0.21	0.0001	0.0073
GO:0000812	Swr1 complex	2	0.05	0.0004	0.0264
GO:0008540	proteasome regulatory particle, base subcomplex	3	0.17	0.0006	0.0330

GO:0044613	nuclear pore central transport channel	2	0.05	0.0006	0.0330
	Molecular Function				
GO:0000166	nucleotide binding	41	16.93	$<10^{-4}$	$<10^{-4}$
GO:0003824	catalytic activity	32	6.88	$<10^{-4}$	$<10^{-4}$
GO:0004069	L-aspartate:2-oxoglutarate aminotransferase activity	2	0.03	$<10^{-4}$	$<10^{-4}$
GO:0004352	glutamate dehydrogenase (NAD+) activity	2	0.03	$<10^{-4}$	$<10^{-4}$
GO:0005524	ATP binding	44	13.51	$<10^{-4}$	$<10^{-4}$
GO:0016491	oxidoreductase activity	22	6.78	$<10^{-4}$	$<10^{-4}$
GO:0016874	ligase activity	9	1.46	$<10^{-4}$	$<10^{-4}$
GO:0030170	pyridoxal phosphate binding	6	0.67	$<10^{-4}$	$<10^{-4}$
GO:0044183	protein binding involved in protein folding	5	0.31	$<10^{-4}$	$<10^{-4}$
GO:0051082	unfolded protein binding	12	0.81	$<10^{-4}$	$<10^{-4}$
GO:0031072	heat shock protein binding	3	0.08	0.0001	0.0144
GO:0004386	helicase activity	6	1.09	0.0003	0.0340
GO:0008483	transaminase activity	4	0.25	0.0003	0.0340
GO:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	4	0.23	0.0003	0.0340

Table S2.3 GO enrichment analysis for antagonism genes

We used a randomization procedure to test whether this group of genes is enriched for GO terms of biological process, cellular component and molecular function. For each GO term, we generated a set of 10,000 random groups of size *N* (where *N* is the number of antagonism genes) sampled from a set that contains antagonism genes and its corresponding background set of genes. In each randomization we computed the number of genes associated to the GO term being tested and used the distribution of the counts across randomizations to calculate the one-tail *P*-values. Only terms overrepresented among antagonism genes after *FDR* correction are shown.

GOID	GO Term	Obs	Exp	*P*	*FDR P*
	Biological Process				
GO:0008299	isoprenoid biosynthetic process	9	2.22	$<10^{-4}$	$<10^{-4}$
GO:0008150	biological_process	382	300.70	$<10^{-4}$	$<10^{-4}$

	Cellular Component				
GO:0005856	Cytoskeleton	42	22.85	$<10^{-4}$	$<10^{-4}$
GO:0005938	cell cortex	34	16.96	$<10^{-4}$	$<10^{-4}$
GO:0016020	Membrane	473	400.65	$<10^{-4}$	$<10^{-4}$
GO:0016021	integral component of membrane	427	352.00	$<10^{-4}$	$<10^{-4}$
GO:0005576	extracellular region	57	30.58	0.0001	0.0082
GO:0005615	extracellular space	87	59.78	0.0001	0.0082
GO:0042995	cell projection	12	3.97	0.0001	0.0082
GO:0005575	cellular_component	441	363.13	$<10^{-4}$	$<10^{-4}$
	Molecular Function				
GO:0003779	actin binding	42	19.60	$<10^{-4}$	$<10^{-4}$
GO:0005515	protein binding	67	42.59	$<10^{-4}$	$<10^{-4}$
GO:0003674	molecular_function	375	307.12	$<10^{-4}$	$<10^{-4}$

Table S2.4 GO enrichment analysis for cheater genes

We used a randomization procedure to test whether this group of genes is enriched for GO terms of biological process, cellular component and molecular function. For each GO term, we generated a set of 10,000 random groups of size *N* (where *N* is the number of cheater genes) sampled from a set that contains cheater genes and its corresponding background set of genes. In each randomization we computed the number of genes associated to the GO term being tested and used the distribution of the counts across randomizations to calculate the one-tail *P*-values. Only terms overrepresented among cheater genes after *FDR* correction are shown.

GOID	GO Term	Obs	Exp	*P*	*FDR P*
	Biological Process				
GO:0035176	social behaviour	24	0.23	$<10^{-4}$	$<10^{-4}$
	Cellular Component				
GO:0005575	cellular_component	35	17.55	$<10^{-4}$	$<10^{-4}$
	Molecular Function				
GO:0016301	kinase activity	11	2.58	$<10^{-4}$	$<10^{-4}$

Table S2.5 Average number of SNPs (SNP/site) for social genes

Expected values and the respective two-tailed *P*-values were obtained from randomization distributions. For each group of social genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance. Significant *p*-values after *FDR* correction for multiple tests are highlighted in bold ($P < 0.05$).

Sites	Group	Expected ($x10^{-3}$)	Observed ($x10^{-3}$)	*P* (*FDR*)
CDS	Sociality	4.758	7.310	< **0.0014**
	Chimerism	4.647	4.321	0.961
	Antagonism	5.252	5.294	0.961
	Prespore	5.252	5.388	0.961
	Prestalk	5.250	5.209	0.961
	Presp-Prest	2.809×10^{-3}	1.784×10^{-1}	0.961
	Cheater	5.087	4.924	0.961
Non-synonymous	Sociality	3.002	4.999	< **0.0014**
	Chimerism	2.861	2.354	0.288
	Antagonism	3.395	3.368	0.961
	Prespore	3.393	3.359	0.961
	Prestalk	3.397	3.376	0.961
	Presp-Prest	-3.563×10^{-3}	-1.703×10^{-2}	0.961
	Cheater	3.280	3.282	0.961
Synonymous	Sociality	1.756	2.311	< **0.0014**
	Chimerism	1.780	1.967	0.560
	Antagonism	1.856	1.926	0.551
	Prespore	1.857	2.029	0.184
	Prestalk	1.858	1.833	0.961
	Presp-Prest	-2.918×10^{-4}	1.954×10^{-1}	0.288
	Cheater	1.799	1.642	0.961

Table S2.6 Complementary neutrality tests for social genes

Fu & Li's statistics compare external and internal branches of a genealogical tree. Under circumstances were variation is removed (purifying selection or recent selective sweeps), it is expected an excess of mutations in external branches (mutations segregating at low frequencies), resulting in negative values. Conversely, balancing selection maintains old alleles (inflating mutations in internal branches), resulting in positive values. Wall's *B* and *Q* statistics use linkage disequilibrium information to test whether a pair of segregating sites share the same genealogy – which would be inflated (larger values) under balancing selection. Expected values and the respective two-tailed *P*-values were obtained by a randomization process. For each group of social genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance after *FDR* correction for multiple tests

Test	Group	Expected	Observed	*P* (*FDR*)
Fu & Li's	Sociality	−0.703	−0.707	0.898
	Chimerism	−0.708	−0.703	0.957
	Antagonism	−0.700	−0.722	0.630
	Prespore	−0.701	−0.762	0.490
	Prestalk	−0.700	−0.685	0.815
	Presp-Prest	0.000	−0.077	0.545
	Cheater	−0.700	−0.850	0.545
Fu & Li's	Sociality	−0.611	−0.602	0.857
	Chimerism	−0.617	−0.594	0.878
	Antagonism	−0.608	−0.623	0.703
	Prespore	−0.608	−0.657	0.545
	Prestalk	−0.607	−0.592	0.824
	Presp-Prest	−0.001	−0.064	0.545
	Cheater	−0.606	−0.717	0.630
Wall's *B*	Sociality	0.087	0.094	0.490
	Chimerism	0.085	0.072	0.545
	Antagonism	0.089	0.082	0.482
	Prespore	0.089	0.078	0.482
	Prestalk	0.090	0.085	0.647
	Presp-Prest	0.000	−0.007	0.642
	Cheater	0.089	0.112	0.545
Wall's *Q*	Sociality	0.116	0.124	0.490
	Chimerism	0.115	0.094	0.545
	Antagonism	0.119	0.109	0.482
	Prespore	0.119	0.105	0.482
	Prestalk	0.119	0.114	0.643
	Presp-Prest	0.000	−0.009	0.630
	Cheater	0.119	0.145	0.545

Table S2.7 Enrichment analysis of social genes evolving under balancing selection as defined by different cutoffs of Tajima's *D*

We used a randomization procedure to test whether each of the groups of social genes contained an excess of genes evolving under balancing selection. For each group of social genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. In each randomization we counted the number of genes evolving under balancing selection and used the distribution of the counts across randomizations to calculate the confidence intervals (2.5^{th} to 97.5^{th} percentiles).

Sites	Group	Observed	CI		*P* (*FDR*)
Tajima's *D* > 2					
CDS	Sociality	13	5	16	> 0.05
	Chimerism	1	0	4	> 0.05
	Antagonism	12	7	20	> 0.05
	Prespore	5	2	12	> 0.05
	Prestalk	7	3	12	> 0.05
	Cheater	1	0	3	> 0.05
Nonsynonymous	Sociality	14	5	16	> 0.05
	Chimerism	1	0	4	> 0.05
	Antagonism	7	7	20	> 0.05
	Prespore	2	2	12	> 0.05
	Prestalk	5	3	12	> 0.05
	Cheater	1	0	3	> 0.05
Synonymous	Sociality	11	5	16	> 0.05
	Chimerism	0	0	4	> 0.05
	Antagonism	12	8	21	> 0.05
	Prespore	8	3	12	> 0.05
	Prestalk	4	2	12	> 0.05
	Cheater	1	0	3	> 0.05
Tajima's *D* > 1.5					
CDS	Sociality	40	26	47	> 0.05
	Chimerism	2	1	9	> 0.05
	Antagonism	47	36	60	> 0.05
	Prespore	22	15	32	> 0.05
	Prestalk	25	16	34	> 0.05
	Cheater	1	0	6	> 0.05
Nonsynonymous	Sociality	40	28	50	> 0.05
	Chimerism	3	1	10	> 0.05
	Antagonism	42	38	62	> 0.05
	Prespore	14	15	33	> 0.05
	Prestalk	28	17	35	> 0.05
	Cheater	1	0	6	> 0.05
Synonymous	Sociality	51	36	60	> 0.05
	Chimerism	8	3	13	> 0.05
	Antagonism	68	52	80	> 0.05

	Prespore	32	23	44	> 0.05
	Prestalk	36	23	44	> 0.05
	Cheater	1	0	7	> 0.05
	Tajima's *D* > 1				
CDS	Sociality	72	61	93	> 0.05
	Chimerism	6	5	17	> 0.05
	Antagonism	94	83	117	> 0.05
	Prespore	45	36	61	> 0.05
	Prestalk	49	39	66	> 0.05
	Cheater	1	1	9	> 0.05
Nonsynonymous	Sociality	71	62	93	> 0.05
	Chimerism	10	5	17	> 0.05
	Antagonism	91	81	116	> 0.05
	Prespore	34	35	60	> 0.05
	Prestalk	57	38	64	> 0.05
	Cheater	3	1	10	> 0.05
Synonymous	Sociality	111	76	109	> 0.05
	Chimerism	13	8	22	> 0.05
	Antagonism	131	108	146	> 0.05
	Prespore	62	50	78	> 0.05
	Prestalk	69	49	77	> 0.05
	Cheater	3	2	11	> 0.05

Table S2.8 Intraspecific variation in sociality genes excluding 13 genes evolving under balancing selection

Expected values and the respective two-tailed *P*-values were obtained from randomization distributions. We generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that sociality genes and its corresponding background set of genes. Significant *P*-values after *FDR* correction for multiple tests are highlighted in bold (*FDR* < 0.05).

Sites	Estimator	Expected ($x10^{-3}$)	Observed ($x10^{-3}$)	*P* (*FDR*)
CDS	π/site	0.770	1.152	< **10^{-3}**
	SNP/site	4.743	7.231	< **10^{-3}**
Nonsynonymous	π/site	0.477	0.772	< **10^{-3}**
	SNP/site	2.992	4.952	< **10^{-3}**
Synonymous	π/site	0.289	0.378	< **10^{-3}**
	SNP/site	1.752	2.278	< **10^{-3}**

Table S2.9 Enrichment analysis of social genes showing strong signatures of selection

We used a randomization procedure to test whether each of the five groups of social genes contained an excess of genes from these two categories. For each group of social genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. In each randomization we counted the number of genes evolving under these forms of selection and used the distribution of the counts across randomizations to calculate the confidence intervals (2.5^{th} to 97.5^{th} percentiles). Significant *P*-values after *FDR* correction for multiple tests are highlighted in bold ($P < 0.05$).

Type of selection	Group	Observed	CI		*P* (*FDR*)
Purifying/Balancing	Sociality	13	6	18	> 0.05
	Chimerism	0	0	3	> 0.05
	Antagonism	10	4	15	> 0.05
	Prespore	8	1	9	> 0.05
	Prestalk	8	1	9	> 0.05
	Cheater	2	0	2	> 0.05
Positive	Sociality	1	2	11	**0.031**
	Chimerism	2	0	3	> 0.05
	Antagonism	9	3	13	> 0.05
	Prespore	6	1	8	> 0.05
	Prestalk	6	1	8	> 0.05
	Cheater	1	0	2	> 0.05

Table S2.10 Evolutionary statistics for prespore and prestalk genes

Expected values and the respective two-tailed *P*-values were obtained from randomization distributions. For each group of genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of prespore or prestalk genes and its corresponding background set of genes. Two-tailed *P*-values are defined as the probability of obtaining a mean as extreme as the observed only due to chance. Significant *P*-values after familywise *FDR* correction for multiple tests are highlighted in bold (*FDR* < 0.05).

Parameter	Group	Expected	Observed	*P* (*FDR*)
π/site	Prespore	0.853 x10^{-3}	0.866 x10^{-3}	0.945
	Prestalk	0.855 x10^{-3}	0.872 x10^{-3}	0.945
	Presp-Prest	-0.001 x10^{-3}	-0.006 x10^{-3}	0.945
π_a/site	Prespore	0.543 x10^{-3}	0.530 x10^{-3}	0.945
	Prestalk	0.542 x10^{-3}	0.571 x10^{-3}	0.919
	Presp-Prest	0.001 x10^{-3}	-0.041 x10^{-3}	0.919
π_s/site	Prespore	0.306 x10^{-3}	0.339	0.616
	Prestalk	0.306 x10^{-3}	0.303	0.945
	Presp-Prest	0.000 x10^{-3}	0.036	0.616
Tajima's *D*	Prespore	-0.628	-0.683	0.214
	Prestalk	-0.628	-0.619	0.738
	Presp-Prest	0.000	-0.064	0.263
Tajima's *D*	Prespore	-0.615	-0.696	0.108
	Prestalk	-0.614	-0.599	0.698
	Presp-Prest	-0.001	-0.097	0.130
Tajima's *D*	Prespore	-0.453	-0.470	0.698
	Prestalk	-0.452	-0.430	0.698
	Presp-Prest	-0.001	-0.040	0.698
DoS	Prespore	-0.016	-0.041	0.638
	Prestalk	-0.017	-0.015	0.901
	Presp-Prest	0.000	-0.026	0.638
$P_n/(P_n+P_s)$	Prespore	0.594	0.591	0.901
	Prestalk	0.594	0.596	0.901
	Presp-Prest	0.000	-0.005	0.901
$D_n/(D_n+D_s)$	Prespore	0.567	0.547	0.638
	Prestalk	0.567	0.564	0.901
	Presp-Prest	0.000	-0.017	0.873
K_a/K_s	Prespore	0.214	0.222	0.722
	Prestalk	0.215	0.169	**0.020**
	Presp-Prest	-0.001	0.053	0.107
K_a	Prespore	0.001	0.001	0.998
	Prestalk	0.001	0.001	0.320
	Presp-Prest	0.000	0.000	0.430
K_s	Prespore	0.009	0.008	0.337
	Prestalk	0.009	0.009	0.722
	Presp-Prest	0.000	-0.001	0.337

Table S2.11 Enrichment analysis of the number of prespore and prestalk genes carrying at least one mutation that introduces a stop codon or results in a partial deletion (presence/absence variation)

We used a randomization procedure to test whether each of the two groups of genes contained an excess of genes carrying these types of deleterious mutations. For each group of genes, we generated a set of 10,000 random groups of size *N* (where *N* is the number of genes in that particular group) sampled from a set that contains that group of social genes and its corresponding background set of genes. In each randomization we counted the number of genes that contained each type of deleterious mutation and used the distribution of the counts across randomizations to calculate the confidence intervals (2.5^{th} to 97.5^{th} percentiles) and *P*-values. Significant *P*-values after *FDR* correction for multiple tests are highlighted in bold ($P < 0.05$).

Class of mutations	Group	Observed	CI		*P* (*FDR*)
Stop codon gain	Prespore	1	1	8	> 0.05
	Prestalk	5	2	10	> 0.05
Presence/Absence	Prespore	0	3	13	**$< 10^{-3}$**
	Prestalk	0	4	14	**$< 10^{-3}$**

3 Processes shaping synonymous codon use in an extremely AT-biased genome

Janaina Lima de Oliveira[1a], Atahualpa Castillo Morales[1a], Laurence D. Hurst[1], Araxi O. Urrutia[1], Christopher R. L. Thompson[2*], Jason B. Wolf[1*]

1. Milner Centre for Evolution and Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK

2. Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London, WC1E 6BT, UK

[a]These authors contributed equally to this work

[*]Correspondence to: jason@evolutionarygenetics.org & christopher.thompson@ucl.ac.uk

Keywords: Codon usage bias; expression optimization; weak selection

3.1 Abstract

Despite their apparent equivalence, synonymous codons are not typically used uniformly in different species and at different genes within the same genome. Deviations from equal synonymous codon usage can reflect selection to optimize expression or simply 'background' processes shaping nucleotide composition. The relative contribution of these adaptive and non-adaptive processes is contentious, with evidence of both processes having been reported. Here, we disentangle the effects of nucleotide composition bias and selection by modelling the expected distributions of synonymous codons under mutation-drift balance in a highly AT-biased eukaryotic genome. We find that mutation bias explains a striking 88% of variation in codon usage bias. Only after accounting for this effect can we identified 'preferred' codons shaped by selection, whose usage increases with expression levels and among genes evolving under stronger selective constraints. Optimization of expression seems to be addressed mostly (but weakly) by shaping levels of transcript stability, addressed by both usage of preferred codons and increasing overall GC content in coding sequences. This pattern suggests a role of selection to counterpoise the strong mutational bias towards AT accumulation in coding sequences. In light of these findings, the need to differentiate 'codon bias' from 'codon preference' is also discussed.

3.2 Introduction

Alternative codons for the same amino acid have been historically regarded as ‘synonymous’, as a result of degeneracy of the genetic code and their presumed interchangeability (Crick et al. 1961; Nirenberg et al. 1966). However, even with scarce sequence data, early comparative studies suggested that synonymous codons are not used in equal frequencies, with biases varying according to general properties of a lineage’s genome (the ‘genome hypothesis’ (Grantham 1980)). Two competing hypotheses have been put forward to explain evolution of synonymous codon use: mutation-drift balance (Kimura 1968; King and Jukes 1969; Sueoka 1988) and natural selection (Ikemura 1985; Akashi and Eyre-Walker 1998; Gingold and Pilpel 2011; Chamary et al. 2006).

Frequently summarized as the proportion of Guanine and Cytosine (GC content), overall base composition is mostly determined by mutational pressures (Sueoka 1962; but see also Rocha and Feil 2010). Although GC content is most often around 50% (with Adenine and Thymine, AT, accounting for the remaining ~50%), strong overall biases are found in both prokaryotes (16.5-75% GC (Muto and Osawa 1987; Nakabachi et al. 2006)) and eukaryotes (19.4-64% GC (Gardner et al. 2002; Merchant et al. 2007)). Supporting the neutral hypothesis of synonymous codon usage as a result of mutation and drift, GC content from the presumably neutrally evolving intergenic regions is a strong predictor of codon usage bias (CUB) across species (Chen et al. 2004). Nucleotide composition may also vary locally within the same genome. The stereotypic example of such phenomenon are the isochores of warm-blooded vertebrates, characterized by large stretches of homogeneous base composition DNA (Bernardi et al. 1985). Similarly to the overall genome scenario, the relative usage of synonymous codons in a gene is largely influenced by base composition of the local region within which the gene finds itself (Bernardi 2000; Urrutia and Hurst 2001).

There is growing evidence that mutations at synonymous sites are not so ‘silent’ and sets of ‘optimal’ codons are actually favoured because are advantageous. Although coding for the same amino acid, alternative codons may

differ in the availability of isoaccepting tRNAs carrying their particular anticodon (Ikemura 1981). This variation can, in turn, affect efficiency (rate) and accuracy (fidelity) of translation (Kurland 1992; Gingold and Pilpel 2011). The arrangement of synonymous codons in a gene can also influence transcript stability, because interactions between base pairs (A:T and G:C) may form secondary structures that increases stability and mRNA steady-state levels (Chamary and Hurst 2005; Wan et al. 2012). More stable transcripts persist longer in the cell, which can result in more translational events per mRNA, increasing the concentration of the final protein (Kudla et al. 2006; Trotta 2013). Therefore, selection to tune expression can act at both translational and transcriptional levels. Moreover, consistent with predictions of evolutionary theory, optimal codons are more broadly used among genes evolving under strong pressures to optimize expression (highly and broadly expressed genes) (Akashi and Eyre-Walker 1998; Akashi 2001) and species with large effective population sizes, such as microorganisms (Ikemura 1985).

In eukaryotes, investigation is often focused on model organisms with ~50% GC, such as mammals (Lander et al. 2001) and flies (dos Santos et al. 2015). This turns the task of quantifying the relative contribution of stochastic processes and selection on CUB particularly challenging. An alternative is to investigate patterns of codon usage bias in organisms with strong nucleotide composition bias, where the presumable effect of background process can be more easily accounted for, before identifying potential signatures of selection.

The free-living social amoeba *Dictyostelium discoideum* has one of the most extreme base composition biases recorded for eukaryotes to date (~22.4% GC (Eichinger et al. 2005)), behind only the human malaria parasite *Paramecium falciparum* (~19.4% GC (Gardner et al. 2002)). Pre-genomic investigation in *D. discoideum* has suggested an influence of base composition in patterns of CUB, with AT-richer synonymous codons being used more frequently but decreasing with expression levels (Sharp and Devine 1989). However, the actual extent to which background processes and selection shapes synonymous codon usage remains unknown.

Here, we integrate large scale genomic and expression data to investigate patterns of CUB and examine the evolutionary processes that have shaped it. By analysing patterns of nucleotide substitution (transitions, transversions and inter-GC class), we first estimated the expected nucleotide composition under mutational equilibrium and used this information to model synonymous codon probabilities under mutation-drift balance. After accounting for a large fraction of variation in CUB explained by neutral processes, we were able to identify sets of 'preferred' and 'unpreferred' codons and analyse their usage in context of gene expression and different selective constraints. Our results show that, although CUB mostly emerges in a passive manner as a result of background processes shaping overall base composition, there is also weak selection to optimize expression features. Usage of optimal codons modulates transcript stability, which can be an important form of optimization of expression in AT-biased genomes, where AT accumulation presumably decreases transcript stability and steady-state levels.

3.3 Results

Evolutionary forces shaping absolute codon usage (i.e. all codons) can act both at the level of nucleotide and protein sequence. Selection favouring protein sequence shapes the use of amino acids, rather than codons, and may be influenced by processes such as selection to reduce biosynthetic costs (Akashi and Gojobori 2002). Selection at the level of amino acid use can potentially blur signatures of evolutionary processes shaping codon usage, particularly given that proteins usually differ in amino acid content. To understand how selection shapes codon use, rather than protein sequences, we focus on processes involved with differential usage of alternative (synonymous) codons for the same amino acid. Consequently, unless otherwise stated, we exclude Methionine and Tryptophan because they only have one codon, and stop codons because they are not directly comparable to amino acids, since their usage is constrained and also cannot be evaluated in terms of some translational optimization processes (such as use of abundant tRNAs). We consider the evolution of protein sequences as a separate (but related) process and analyse processes shaping protein evolution elsewhere (de Oliveira et al. in prep.a).

3.3.1 AT-richer codons are used more frequently

To investigate the potential influence of background nucleotide composition on synonymous codon usage, we assessed the observed relative frequency of synonymous codons and their GC content. Remarkably, the more frequent codon for every amino acid and stop signal is always one of the AT-richest alternatives (Figure 3.1A). This pattern is even stronger when alternative codons with the same GC content are considered as a single category: the AT-richer codons are used up to 96% of the time, with an average use of 86% across all amino acids (Figure 3.1B).

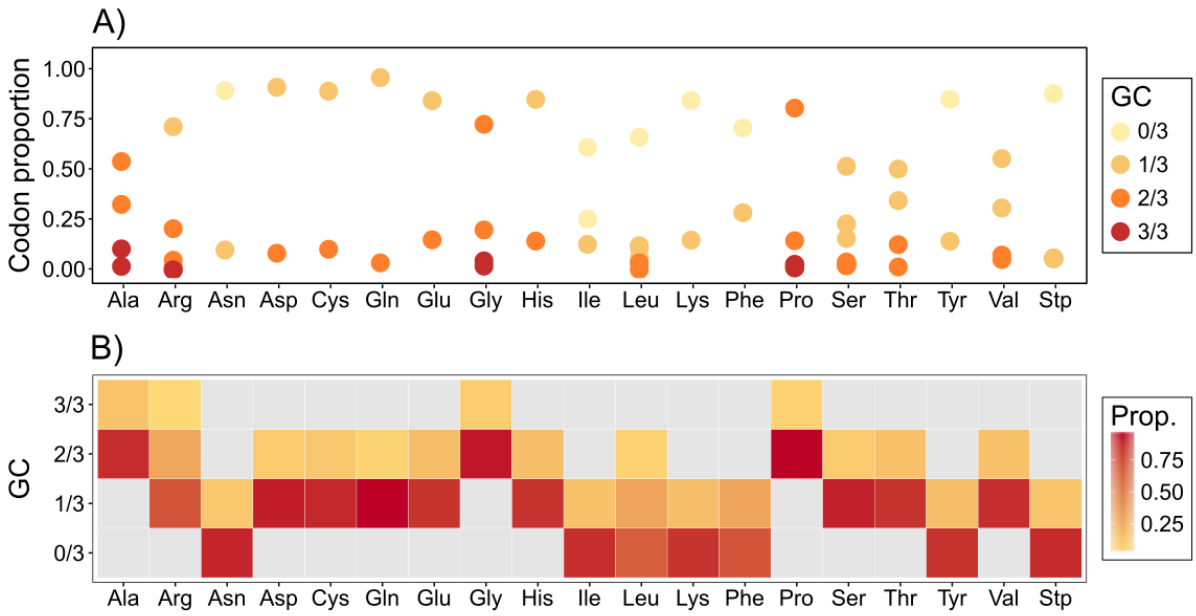

Figure 3.1 Relative codon frequencies and GC content.

A) Proportional use of each codon for each amino acid and stop signal, with points representing the individual codons and the colours their GC content. **B)** The pooled proportional use of codons per amino acid grouped by their GC content.

Because codon usage can be influenced not only by overall base composition in the genome, but also by the local composition around a given gene (Urrutia and Hurst 2001), we performed a sliding window analysis to assess the distribution of GC across the genome. Apart from a few peaks that appear to be related to an enrichment of transposable elements (TEs), GC content is evenly distributed across all chromosomes (Figure S3.1). Moreover, processes underlying

base composition in surrounding non-coding regions (introns and intergenic regions) explain only a small fraction of GC in coding regions (R^2 = 0.0189, P < 0.0002). This weak correlation is no longer significant after removing outlier peaks of GC due to an overrepresentation of TEs in chromosomes 1 (bases 1-200Kb) and 6 (bases 850-900Kb) (R^2 = 0.0023, P < 0.1105). These results suggest that processes shaping overall, rather than local, base composition influence GC content and synonymous codon usage in coding sequences.

3.3.2 Mutation bias explains a large proportion of synonymous codon use

To quantify the extent to which synonymous codon usage simply reflects background processes shaping nucleotide content in the genome, we model the expected distribution of alternative codons under the null hypothesis that it is driven by base composition under equilibrium (GC_{eq}). The equilibrium base composition can be modelled using direct estimates of mutation rates obtained, for example, from mutation accumulation experiments, or from indirect estimates such as those based on rare segregating variants from regions evolving under neutrality or close to neutrality. Although previous work on mutation accumulation lines have investigated mutational patterns in this system (Saxer et al. 2012), conclusions were drawn from a single mutation that emerged throughout the experiment, and thus provides only a very rough approximation of the mutation rate and no measure of the differential mutation rate between the four nucleotides. Here we use SNP data from non-coding regions of 67 natural strains (de Oliveira et al. in review) to extract information about the underlying mutation process and compute the nucleotide substitution matrix (Figure 3.2; Figure S3.2).

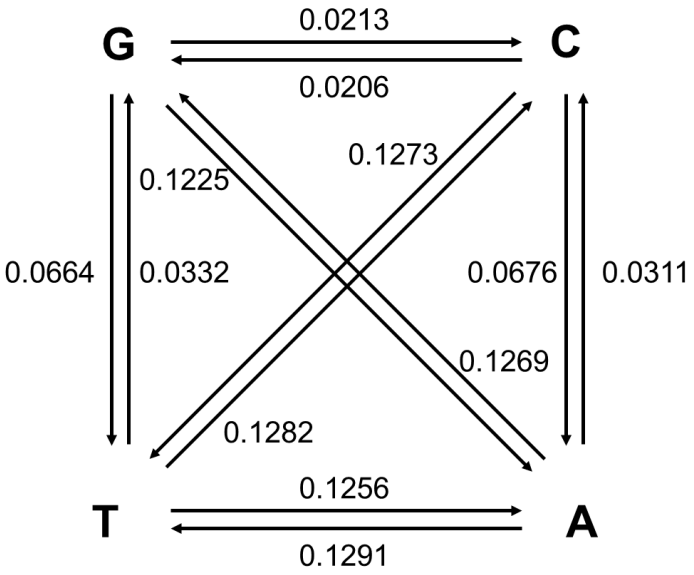

Figure 3.2 Nucleotide substitution matrix.

For each SNP, variants where classified as ancestral (the nucleotide segregating at higher frequency) and derived (the nucleotide segregating at lower frequency). Derived variants were considered mutations from the ancestral allele, and the proportion of all mutations from one nucleotide to each other nucleotide was estimated. Values are proportions of mutations that belong to each category. For example, 3.11% of all mutations are A to C transversions.

The analysis of mutations reveals that mutations at an A or T have roughly the same chance of being a transition (A → G and T → C) as they do of being a transversion towards the alternative AT nucleotide (A → T and T → A), while mutations only rarely constitute a transversion away from AT (A → C and T → G). In contrast, mutations at a G or C are primarily transitions towards AT (G → A and C → T), followed by transversions towards AT (G → T and C → A), and only rarely do mutations oppose this flow, staying in the same GC class (G → C and C → G). Although transitions (which always change GC class) towards and away from AT occur roughly in the same proportion, transversions towards AT are more common than towards GC. More remarkably, intra-class mutations are much more common in the AT than in the GC category. An intuitive outcome of this pattern is a ‘loss’ of GC content, mostly because mutations at G and C rarely conserves the nucleotide in the GC category, whereas mutation in A and T often conserves the nucleotide in the AT category.

To predict the equilibrium GC content of the genome (GC_{eq}) under our estimated pattern of mutation we extended a previous model (Sueoka 1962) to account for the proportion of mutations that retain the same GC class (A → T, T → A, G → C, C → G). This extension is necessary because the probability of staying in the same GC class differs between GC and AT pairs (Figure 3.3), influencing equilibrium estimates. By applying this method, we predict GC_{eq} to be around 16%, which is remarkably close to the observed in non-coding regions (14%). This figure is at odds with the considerably higher GC content in coding regions (~27%), which may reflect a selective process opposing the strong bias towards AT accumulation in coding sequences.

If usage of alternative codons is random and driven by background processes shaping base composition, we would expect observed relative synonymous codon frequencies to be very close to that expected under GC_{eq}. To test for this possibility, we modelled expected synonymous codon frequencies under neutrality using a scenario in which codon use is simply a product of base composition probabilities at each position of a codon, rescaled for each amino acid (see Methods). This expected distribution of synonymous relative frequencies explains a remarkable 88% of variation in observed codon frequencies ($R^2 = 0.88$, $P < 10^{-15}$; Figure 3.3A).

The random codon use model may be an oversimplification of the neutral expectation if complex patterns of trinucleotide mutations occur, making the probability of a given triplet more than the simple 'sum of its parts'. Such a scenario can potentially be more likely at repeat-rich DNA sequences, because the repeats increase the chances of polymerase slippage during DNA replication (Ellegren 2004). Because *D. discoideum* has a repeat-rich genome (Eichinger et al. 2005), we tested whether the random occurrence of triplets deviates from that predicted from the product of nucleotide frequencies. Occurrence of triplets under neutrality was estimated by computing the number of triplets in non-coding regions in all possible three frames. These counts were then treated as regular codons to obtain relative synonymous codon frequencies, should these triplets be actually translated into amino acids. A linear regression analysis reveals that relative synonymous codon frequencies of these neutrally evolving triplets is strongly predicted from nucleotide

composition under equilibrium (R^2 = 0.92, $P < 10^{-15}$), suggesting that base frequencies alone can be used to estimate expected synonymous codon distribution under neutrality.

We also considered the alternative hypothesis that observed relative frequencies can be best explained by coevolution between codons and available isoaccepting tRNAs (Ikemura 1981; Ikemura 1985). Because tRNA gene copy number is strongly and positively correlated to tRNA levels in a cell, it can be used as a reliable proxy of the later (dos Reis et al 2003). dos Reis et al.'s (2003; 2004) relative codon adaptiveness (w_i) is defined as the number of tRNA copies that recognize a particular codon after accounting for wobble pairings – the absolute codon adaptiveness (W_i) – weighted by the maximum absolute codon adaptiveness among all codons (W_{max}). However, because our interest is on usage of alternative codons for the same amino acid (and w_i is defined relatively to all codons), we derived a new parameter: the relative synonymous codon adaptiveness, w_{ij}. Here, the absolute adaptiveness of a codon *i* is defined in context of amino acid *j*, and weighted by the maximum absolute adaptiveness among codons for amino acid *j* (W_{jmax}) (see Methods). We find that only a small fraction of observed relative frequencies is explained by w_{ij} (R^2 = 0.14, $P < 0.002$; Figure 3.3B).

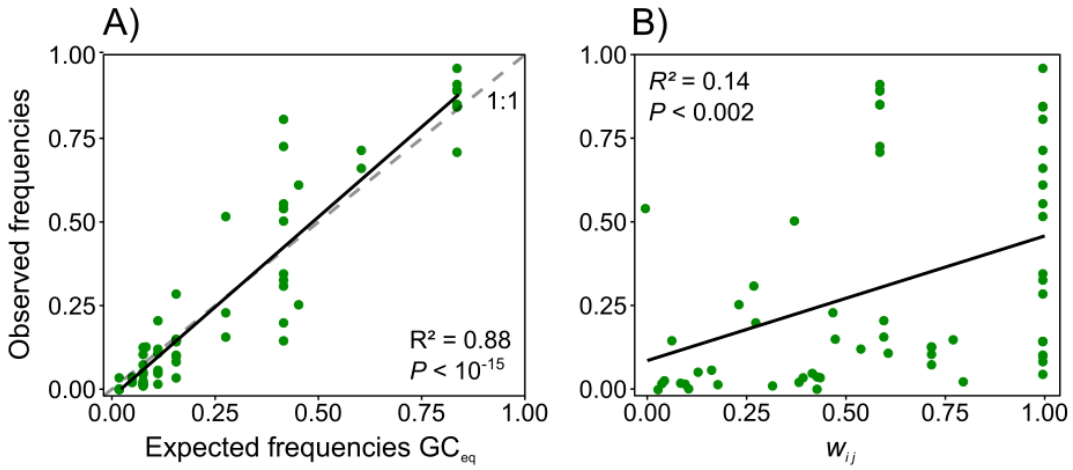

Figure 3.3 Contribution of neutral and adaptive processes to observed synonymous codon frequencies.

A) Observed frequencies can be strongly predicted by the distribution expected under neutrality, derived from base frequencies under GC_{eq}. **B)** Analysed as a continuous variable, synonymous relative codon adaptiveness (w_{ij}) explains only a fraction of observed frequencies.

An extended model to explain observed relative frequencies that includes both expected frequencies under neutrality and w_{ij}, although significantly better (in comparison to the model that includes only expected frequencies, ANOVA $P < 0.004$), explains only a marginal extra 1% of the variation ($R^2 = 0.89$, $P < 10^{-15}$). These results suggest that a large proportion of the observed frequencies that is explained by w_{ij} is also explained by background processes, with biases in synonymous codon frequencies mostly arising from a passive process, under the influence of random forces shaping base composition towards AT accumulation.

3.3.3 CUB is also shaped by selection to optimize expression

A strong influence of base composition in CUB does not necessarily reflect an absence of selection, but suggests that putative signatures of selection can be obscured by the large impact of neutral processes. Moreover, the potential impact of neutral and adaptive processes shaping CUB can be widely diverse across individual genes. 'Preferred' and 'unpreferred' codons can be identified by over

and under usage compared to the neutral expectation (this nomenclature is adopted to avoid confusion with 'optimal' and 'non-optimal' codons arising from co-evolution with tRNA pools). If these deviations reflect differential fitness of alternative codons on transcriptional/translational features, we would expect two patterns. First, the relative usage of preferred codons should increase with expression, since genes required at higher and broader expression evolve under stronger selective pressures. Second, genes evolving under different selective constraints would show differences on their distribution of preferred codons to optimize expression.

We used a collection of publicly available transcriptomes of the vegetative and developmental cycles of *D. discoideum* (Nasser et al. 2013; Parikh et al. 2010; Rosengarten et al. 2015) to obtain gene expression levels, and found a positive correlation between maximum expression and usage of preferred codons ($r = 0.40$, $P < 10^{-15}$; Figure 3.4A). Moreover, we compared these patterns between two groups of genes (Sociality and Non-sociality) previously identified to evolve under different selective constraints (de Oliveira et al. in review). Evolution at the set of Sociality genes reflects the Red King process, where the strength of selection is diluted due to conditional expression (as a result of expression being restricted to the social cycle), whereas genes expressed in every generation (Non-sociality genes) do not show this signature. Interestingly, we found that, across the whole coding sequence, Sociality genes tend to show lower usage of preferred synonymous codons in comparison to Non-sociality genes (Figure 3.4B). Furthermore, both Sociality and Non-sociality genes show a positive correlation between usage of preferred codons and expression (Non-sociality: $r = 0.43$, $P < 10^{-15}$; Sociality: $r = 0.38$; $P < 10^{-15}$; Figure 4C), but the latter group shows a significantly weaker relationship ($z_{\text{Non-sociality}} - z_{\text{Sociality}} = 2.27$, $P = 0.023$). These findings suggest that selection plays an important role in shaping CUB in this system in order to optimize expression.

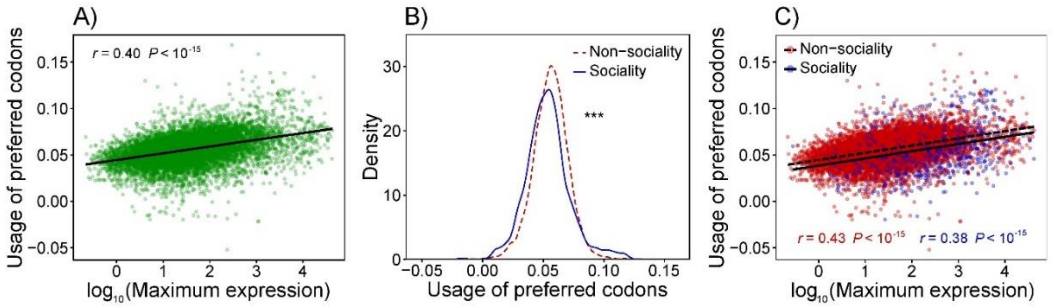

Figure 3.4 Patterns of inferred selection on codon usage bias.

A) Usage of preferred codons increases with expression levels, as expected under translational/transcriptional selection. **B)** Genes with conditional expression (Sociality genes), which evolve under diluted selective constraints, show a reduced overall usage of preferred codons compared to genes that experience selection in all generations (Non-sociality genes). **C)** Codon usage bias and expression levels are positively correlated in both groups of genes evolving under different selective constraints, but this correlation is weaker in genes evolving under diluted selection.

3.3.4. Expression optimization is accomplished by modulating transcript stability

Optimization of expression is often achieved by coevolution with tRNA availability, but we have shown that relative synonymous codon adaptiveness is very weakly correlated to relative synonymous codon frequencies. Yet, we see a positive correlation between codon bias and expression levels, and with stronger selective constraints (Figures 3.4A and 3.4C). So, which mechanism explains the apparently adaptive relationship between usage of preferred codons and expression? One possibility is that this is indeed achieved by co-evolution with tRNA abundance, but that this signature was not captured by our initial analysis because it can be only revealed when analysed at individual transcripts. A second possibility is that selection acts in another form, such as by increasing stability and mRNA steady-state levels (Trotta 2013; Chamary and Hurst 2005), which we refer to as transcriptional selection.

To gain insights as to why some codons are preferred (favoured by selection), we integrate expression levels with factors potentially related to translational and transcriptional selection, and investigate their influence on

differential usage of synonymous codons. As a mechanism of translational selection, we defined a gene's synonymous tRNA adaptation index (StAI), adapted from dos Reis et al's (2003; 2004) tRNA adaptation index (tAI). Accordingly, StAI is intended to measure the relationship between overall codon usage in a gene and the available tRNA pool (using tRNA copy numbers as a proxy). However, StAI is calculated as the geometric mean of w_{ij} (codon fitness relative to synonymous codons for the same amino acid) across a gene, not w_i (codon fitness relative to all codons; see Methods). As a mechanism of transcriptional selection, we estimated a measure of stability per site (for short, hereafter referred simply as 'stability'), defined as the opposite (negative) of Gibb's free energy (i.e. $-\Delta G^o$) divided by transcript length.

Consistent with the analysis of individual codons, the usage of preferred/unpreferred codons (codon preference) by a gene is very weakly correlated with usage of synonymous codons with more available isoaccepting tRNAs (StAI) (r = 0.05, $P < 10^{-5}$; Figure 3.5A). Conversely, usage of preferred/unpreferred codons is positively correlated to stability of a transcript (r = 0.34, $P < 10^{-15}$; Figure 3.5B). These findings suggest that although weak, selection to optimize expression by usage of differential codons is mostly achieved at the transcript level by increasing transcript stability.

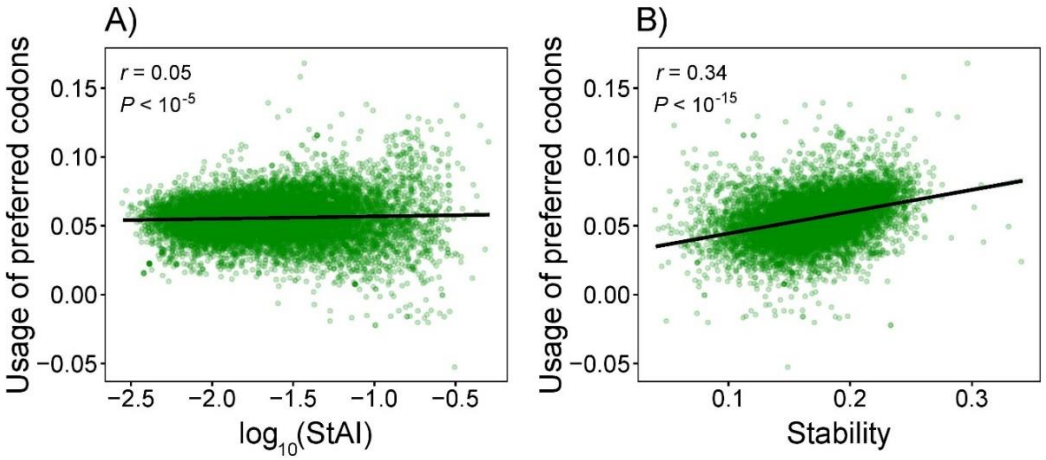

Figure 3.5 Relationship between codon preference and expression optimization parameters.

A) Correlation between codon bias and usage of synonymous codons by isoaccepting tRNA availability (StAI). **B)** Correlation between codon usage bias and transcript stability ($-\Delta G^o$).

Preferred codons identified by our analysis are corrected for GC content bias, since the underlying mutational process is taken into account in the calculation of expected synonymous codon frequencies. However, we hypothesized that, if transcript stability is a general trait under selection to optimize expression, then selection could also shape base composition in coding sequences by increasing overall GC content, since G:C bonds are more stable than A:T. The GC content of a coding sequence is indeed positively correlated with transcript stability ($r = 0.57$, $P < 10^{-15}$; Figure S3.3). We further analysed the GC content of a coding sequence to test for an effect of expression and a difference between groups evolving under different selective constraints (Figure 3.6) following an approach that is similar to our analysis of codon bias and expression (Figure 3.4). As predicted, we find that GC content shows a very strong positive relationship with expression level ($r = 0.68$, $P < 10^{-15}$; Figure 3.6A) and is slightly higher in genes evolving under stronger selective constraints (Average $GC_{Non\text{-}sociality} = 0.28$, Average $GC_{Sociality} = 0.27$, $P < 10^{-15}$; Figure 3.6B). Moreover, while GC content increases with expression in both Sociality ($r = 0.67$, $P < 10^{-15}$; Figure 3.6C) and Non-sociality genes ($r = 0.71$, $P < 10^{-15}$; Figure 3.6C), the relationship is weaker in the former ($z_{Non\text{-}sociality} - z_{Sociality} = 2.97$, $P = 0.003$).

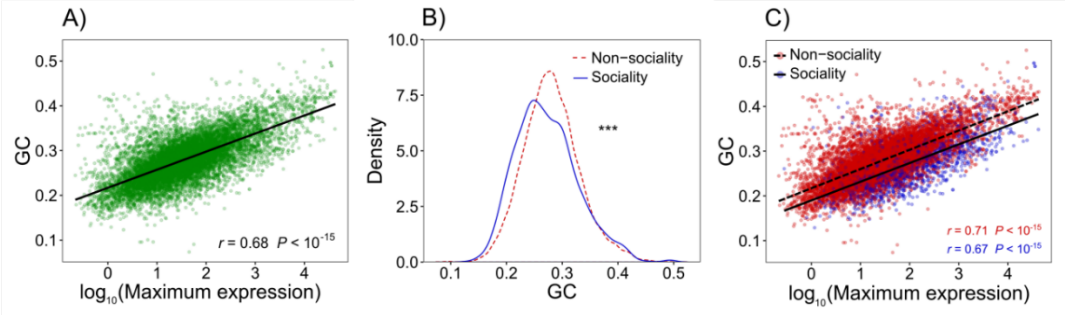

Figure 3.6 Selection on overall GC content in coding regions.

A) GC content is strongly and positively correlated with expression levels. **B)** Genes evolving under diluted selective constraints (Sociality genes) show a shift in the distribution of GC content towards the neutral expectation of low GC. **C)** GC content and expression levels are positively correlated in both groups of genes evolving under different selective constraints, but this correlation is weaker in genes evolving under diluted selection (Sociality genes).

3.4 Discussion

The relative contribution of adaptive and non-adaptive processes is one of the most debated topics in molecular evolution. Accordingly, observation that synonymous codons are not used in equal frequencies has motivated several competing hypotheses, which can be largely separated into two categories: those postulating that codon usage bias emerges passively as a result of background (neutral) processes (Kimura 1968; King and Jukes 1969; Sueoka 1988), and those suggesting an active role of selection favouring codons that optimize expression (Ikemura 1985; Akashi and Eyre-Walker 1998; Trotta 2013). We used a model system with a highly AT-biased genome, the social amoeba *D. discoideum*, to quantify the potential effects of base composition and selection to optimize expression. Analysis of relative synonymous codon frequencies reveals a clear trend towards usage of AT-richer codons (Figure 3.1), as previously indicated by pre-genomic investigation in this system (Sharp and Devine 1989) and reports from the reference genome (Eichinger et al. 2005). A large fraction of codon usage bias can be explained solely by mutational biases towards AT accumulation in this genome (Figure 3.2). Specifically, expected frequencies under neutral evolution

explains a remarkable 88% of variation in observed synonymous codon usage, which is around six times higher than the fraction explained by co-evolution with the tRNA pool (Figure 3.3).

These findings do not exclude the possibility of a role of selection to optimize expression features. Signatures of translational/transcriptional selection are often revealed in the context of gene expression, since broadly and highly expressed genes evolve under stronger constraints to optimize expression (Akashi and Eyre-Walker 1998; Akashi 2001). Moreover, the strong effect of mutational bias can potentially obscure signatures of selection, since the latter is expected to be weak at synonymous codons (Akashi 1995). After removing the strong effect of base composition on synonymous codon bias, usage of preferred codons (used more often than the neutral expectation) increases with expression and among genes evolving under stronger selective constraints (Figure 3.4).

Favouring of preferred codons by selection does not appear to be caused by a strong adaptation to use abundant synonymous tRNAs, but to a weak tuning of transcript stability through modulation of mRNA secondary structure (Figure 3.5) – similarly to results reported from mammals (Chamary and Hurst 2005). Stability increases mRNA steady-state levels, and has been implied as an important mechanism of optimization of expression (Kudla et al. 2006; Trotta 2013; Chamary and Hurst 2005). In eukaryotes, regulation of gene expression by varying levels of mRNA stability involves the proper processing of the transcript by addition of a cap and a poly(A)-tail at the 5' and 3' ends, respectively, which protect the transcript from exonucleases attack (Garneau et al. 2007). Moreover, secondary structures formed at 5' and 3' UTRs have been experimentally demonstrated to modulate levels of gene expression, either by influencing translation initiation (Dvir et al. 2013) or mRNA stability (Moqtaderi et al. 2014). However, regulatory mechanisms at the transcript are not limited to features of untranslated regions and structures added at the transcript ends (cap and poly(A)-tail). In yeasts, it has been experimentally demonstrated that transcript melting temperature (T_m, used as a measure of transcript structure stability) and progressive inclusion of hairpins influence rates of RNA decay by the exosome (Wan et al. 2012), proving evidence for the role of secondary structures in transcript stability.

Considering that *D. discoideum* evolves under a strong trend towards AT accumulation (Figure 3.2) – and A:T forms weaker bonds than G:C – we speculate that mechanisms that increases mRNA stability are favoured by selection. Accordingly, both usage of preferred codons and overall GC content (Figures 3.5 and 3.6), which both increase transcript stability, are positively correlated with expression levels, and more common among genes evolving under stronger selective constraints. Future experimental work comparing stability and expression of alleles coding for the same protein but with different proportions of preferred/unpreferred codons, as well as different overall GC contents, might provide interesting insights on this hypothesis.

Codon bias can emerge by both passive and active processes. In genomes with strong base composition bias, departures from equal usage of synonymous codons can emerge passively, by simply drifting towards the genome-wide base composition. This is the case in *D. discoideum*, where usage of AT-richer codons is strongly influenced by background processes shaping nucleotide composition. However, once this effect is accounted for, an active role of selection is clear, with preferred codons being effectively favoured by selection because they confer a selective advantage at the translational and/or transcriptional level. Because the overall pattern of codon bias emerges from a passive process driven by mutational bias, it does not truly reflect what we would consider 'codon preference', unlike in most systems where commonly used codons are those 'preferred' (i.e. favoured by selection). We see that a pattern of truly preferred codons emerges when we consider how relative use of codons changes in relation to inferred sources of selection. Thus, we suggest that the term 'codon preference' should be reserved for departures of codon usage from the neutral expectation (presumably driven by mutational processes) caused by an active role of selection arising from different codon fitness, whereas 'codon bias' can be used to describe any deviations from equal usage of synonymous codons emerging from either passive or adaptive processes.

3.5 Methods

3.5.1 Synonymous codon frequencies and GC distribution across the genome

Relative synonymous codon frequencies (i.e. relative to each amino acid) were estimated from the reference genome (Eichinger et al. 2005) downloaded from Ensembl (Aken et al. 2016; Kersey et al. 2016) and using the R package seqinR (Charif and Lobry 2007). Before computing this codon table, we excluded all non-protein coding sequences, genes from the mitochondrial genome and from a duplication in chromosome 2, present only in the strain AX4 (reference genome, Eichinger et al. 2005). This censoring was necessary because the codon unity is meaningful only when translated into amino acids, and because genes in these other regions (mitochondrion and duplication) can evolve under different dynamics in comparison to the rest of the genome.

GC content was computed in coding and non-coding regions of all six chromosomes of *D. discoideum*, in windows of 50Kb separated by step sizes of 1Kb. In each window, we used coordinates from Ensembl (Aken et al. 2016; Kersey et al. 2016) to characterize chromosome regions as coding or non-coding DNA. We also used these coordinates to localize a list of genes annotated as transposable elements in Dictybase (Fey et al. 2013), to test the hypothesis that peaks of elevated GC could be associated to the presence of such elements. Peaks of both lower and higher GC ($< 5^{th}$ and $> 95^{th}$ percentiles of GC distribution in the 50Kb windows) were identified from non-coding regions, under the assumption that non-coding DNA evolve close no neutrality (whereas base composition could be potentially under selection in coding sequences).

3.5.2 Nucleotide substitution matrix and GC_{eq}

Overall nucleotide composition is mostly a result of mutational biases (Sueoka 1962; but see also Rocha and Feil 2010), so understanding the evolution of such an AT-biased genome as in *D. discoideum* must include a detailed investigation of mutational processes. Because experimental work on mutational patterns in this system resulted in conclusions drawn from a single SNP (Saxer et

al. 2012), we used information from segregating variation to derive general patterns. This dataset includes 67 natural strains, and details on the geographical distribution of the strains, sequencing reports, mapping and SNP calling are provided elsewhere (de Oliveira et al. in review). SNPs were filtered to include only those from non-coding regions, since these must reflect evolution close to mutation-drift balance. Directionality was inferred from polarization of rare alleles in comparison to the common alleles, resulting in a nucleotide substitution matrix with proportion of substitutions in all directions of mutational space.

This information was used to derive the expected GC under equilibrium. Sueoka's (1962) original equation does not distinct AT → TA and GC → CG. However, because the chances of staying in the same GC class differs between AT and GC categories (see Figure 3.2), and this can presumably affect estimates of equilibrium, we extended his equation as follows:

$$GC_{eq} = \frac{(\mathrm{AT} \rightarrow \mathrm{GC}) - (\mathrm{AT} \rightarrow \mathrm{TA})}{(\mathrm{AT} \rightarrow \mathrm{GC}) - (\mathrm{AT} \rightarrow \mathrm{TA}) - (\mathrm{GC} \rightarrow \mathrm{CG}) + (\mathrm{GC} \rightarrow \mathrm{AT})} \quad (1)$$

3.5.3 Expected relative synonymous frequencies and identification of preferred codons

The estimated value of GC_{eq} (~16%) from equation (1) was first used to calculate the expected absolute codon frequencies, which were later weighted by the expected amino acid frequencies to obtain relative codon frequencies. For instance, consider the codon AAA, one of the two codons (besides AAG) that code the amino acid Lysine. The relative frequency of AAA can be defined as:

$$f(AAA) = \frac{f(A)^3}{f(Lys)} \quad (2)$$

where *f(Lys)* is defined as:

$$f(Lys) = f(AAA) + f(AAG)$$
$$= f(A)^3 + (f(A)^2 \times f(G)) \quad (3)$$

The alternative method of estimating expected relative frequencies by computing the emergence of triplets under neutrality was performed as follows. Non-coding regions of all six chromosomes were concatenated in a single linear sequence, which was divided in triplets on frames 1, 2 and 3. Relative frequency of a triplet was defined as the sum of counts of the triplet in all 3 frames, divided by the total number of triplets for the same 'amino acid' (if they were from coding sequences). Because estimate using this method is very close to the one based on the product of base frequencies, we used the simpler method to calculate expected relative synonymous frequencies.

To account for base composition bias, we identified sets of preferred and unpreferred codons by subtracting the observed relative synonymous codon frequencies from the relative synonymous codon frequencies expected under equilibrium. This method assumes that codons used more often than predicted under neutrality ($Obs_f > Exp_f$) must confer an advantage and are favoured by selection. Conversely, codons used less frequently than expected by neutral evolution ($Obs_f < Exp_f$) are assumed to confer a disadvantage and are therefore unpreferred/avoided. These residuals are averaged across the whole coding sequence, to give a gene's overall index of codon usage preference.

3.5.4 Parameters of translational and transcriptional selection

Co-evolution of codons with the pool of isoaccepting tRNAs is one of the more widespread mechanisms of optimization of expression in nature. One classical analysis to test this hypothesis is based on two related measures: the relative codon adaptiveness (w_i), and a gene's index of tRNA adaptation, tAI (dos Reis, Wernisch, and Savva 2003; dos Reis, Savva, and Wernisch 2004). The first gives a measure of fitness assigned to each codon, whereas the second uses w_i across the whole coding sequence to assign an adaptiveness value for a gene. A limitation of this method is that w_i (and, consequently, tAI) is defined relatively to the maximum

adaptiveness value across all codons (W_{max}), including codons for different amino acids. This means that it may not be an appropriate measure for understanding the usage of alternative codons, particularly at the gene level, since different indices of tRNA adaptiveness may be due to differences in amino acid content rather than differences on the strength of selection on synonymous codons to optimize expression.

To convert this analysis to one that is appropriate for the study of synonymous codons, we re-scaled both the measure of codon adaptiveness and the gene's index of tRNA adaptation to such that they measure the relative values for different synonymous codons for each individual amino acid. Thus, the relative synonymous codon adaptiveness of a particular codon (w_{ij}) is defined as:

$$w_{ij} = \frac{W_i}{W_{jmax}} \quad (4)$$

where W_i is the absolute codon adaptiveness (tRNA gene copy numbers after accounting for wobble pairings), and W_{jmax} is the maximum absolute adaptiveness among codons of amino acid *j*.

The synonymous tRNA adaptation index (StRNA) is defined as the geometric mean of w_{ij} in all positions of a gene:

$$StAI = \left(\prod_{i=1}^{L} w_{ij} \right)^{1/L} \quad (5)$$

where *L* is the sequence length after removing codons for Methionine, Tryptophan (both with a single codon) and stop signal. Both parameters (w_{ij} and StAI) where estimated by adapting R scripts from dos Reis et al (2003; 2004).

As a measure of transcriptional selection, we estimated levels of transcript stability based on Gibbs free energy (ΔG^o), using ViennaRNA package (Lorenz et al. 2011). Given the same transcript length, transcripts with lower (more negative) ΔG^o are more stable. Thus, we divided this measure by CDS length and multiplied

the ratio by −1, converting the original parameter into a weighted and more intuitive measure of transcript stability.

To understand the relationship between overall GC and expression optimization, we also estimated GC content for the whole coding sequence using seqinR (Charif and Lobry 2007).

3.5.5 Expression levels and genes evolving under different selective constraints

Expression levels were defined as the peak of maximum expression after normalization of vegetative and developmental RNAseq libraries (Parikh et al. 2010; Nasser et al. 2013; Rosengarten et al. 2015). Details of the analysis are provided by Oliveira et al (in review) and only briefly outlined here. Libraries were normalized using the TMM method (Robinson and Oshlack 2010) implemented in edgeR (Robinson, et al. 2010), after removing genes with low counts, following author's specifications. Maximum instead of average or breadth of expression was used because the developmental (social) cycle of these species is conditional – i.e. it only occurs if/when the amoebae starve. Thus, defining an expression parameter for genes with a single high peak on late development based in comparison to all libraries may blur putative signatures of selection to optimize expression features, because this gene would presumably have a low average and breadth of expression.

Conditionality of the social cycle has been shown to have an important impact on evolution of the genes expressed only in the social cycle because it dilutes selection, resulting in the Red King process in which genes show signatures that are closer to the neutral expectation compared to non-conditionally expressed genes (de Oliveira et al. in review). These conditionally expressed 'Sociality genes' were compared to 'Non-sociality genes', which are expressed in every generation to test specific hypotheses on the influence of selection under different selective constraints.

3.6 Supplementary material

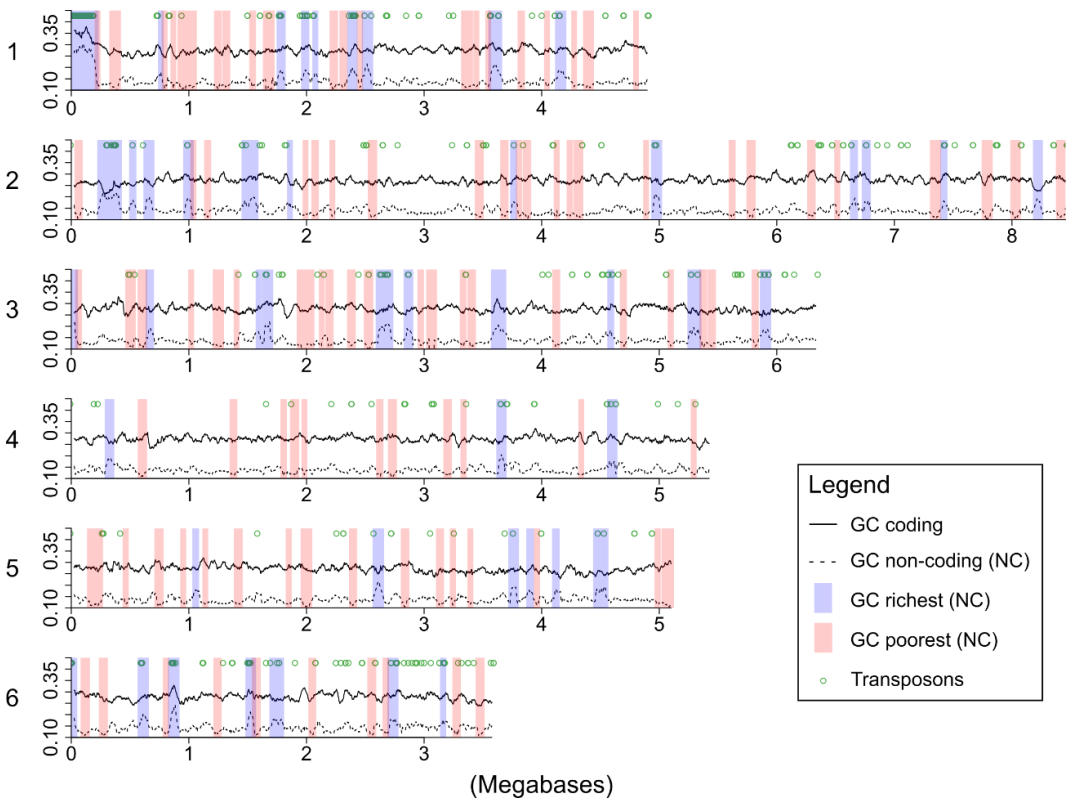

Figure S3.1: Sliding window analysis of genomic GC content.

GC content was estimated for coding (solid lines) and non-coding (dashed lines) sequences in 50Kb windows in 1Kb step sizes across all six chromosomes of *D. discoideum*. Regions with lowest (< 5th percentile) and highest (> 95th percentile) GC contents in non-coding sequences are highlighted in red and blue bars, respectively. Peaks of greatest GC (mostly on chromosomes 1 and 6) are associated with an overrepresentation of transposable elements (green dots).

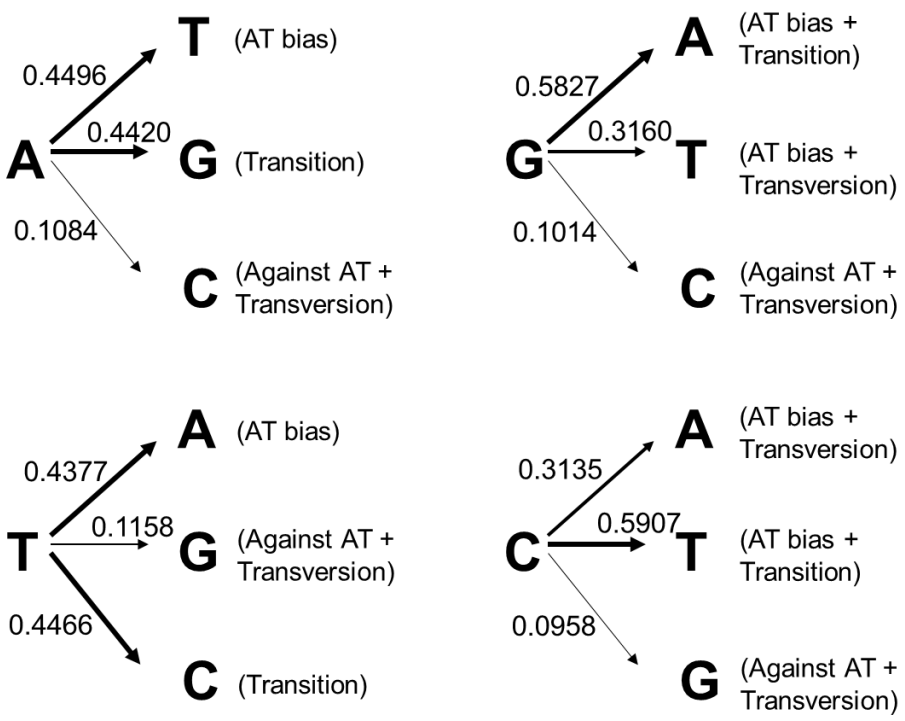

Figure S3.2: Relative nucleotide substitution matrix.

Numbers indicate the fraction of substitutions (minor SNP class) towards each direction of the mutational space. Mutations are categorised as transitions, transversions and either following or against the overall bias towards AT accumulation (see text for more details).

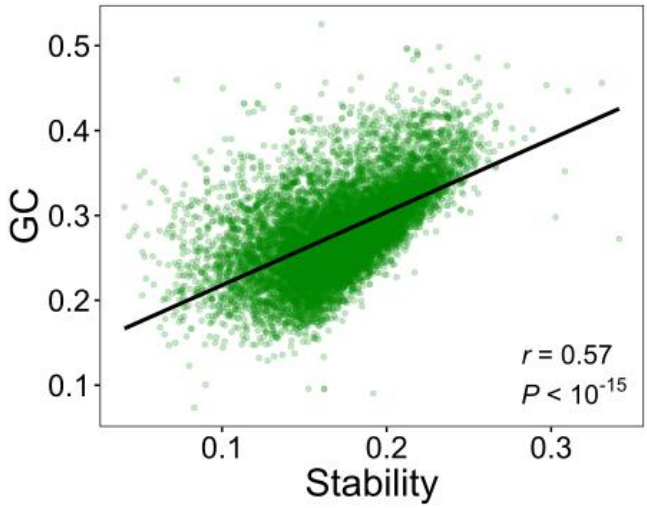

Figure S3.1 Overall GC and transcript stability.

Correlation between overall coding sequence's GC and transcript stability ($-\Delta G^{o}$).

4 Amino acid composition is influenced by evolutionary processes shaping genomic and metabolic features in a microbe

Janaina Lima de Oliveira[1], Atahualpa Castillo Morales[1], Laurence D. Hurst[1], Araxi O. Urrutia[1], Christopher R. L. Thompson[2*], Jason B. Wolf[1*]

1. Milner Centre for Evolution and Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK

2. Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London, WC1E 6BT, UK

[*]Correspondence to: jason@evolutionarygenetics.org & christopher.thompson@ucl.ac.uk

Keywords: Protein evolution; amino acid content; weak selection

4.1 Abstract

Natural selection shapes the sequence of amino acids in proteins to optimize protein function in the face of constraints (e.g., relative costs or availability of different amino acids, transcription efficiency, translation speed etc). These adaptive processes interact with random background processes (mutation and random drift) to yield the observed patterns of amino acid use we observe in the genome. Understanding the relative importance of these adaptive and non-adaptive factors can provide important insights into the composition of proteins. However, in many systems it can be difficult to disentangle their influence. Here we exploit the extremely biased nucleotide composition of the *Dictyostelium discoideum* genome to reveal the relative importance of these factors. We find that mutational bias is the largest driver of amino acid composition, but once accounted for, we uncover the underlying influence of metabolic costs. The impact of mutational bias declines rapidly with the level of gene expression, presumably reflecting the increased importance of protein optimization (with amino acid composition depending on the distinct peculiarities of individual proteins), while the importance of cost minimization increases. These findings highlight the importance of including contextual information on the study of protein evolution, rather than viewing a protein as an isolated entity.

Assessing information from molecular variation was a remarkable step in testing validity of theoretical population genetics models. From the debate over the relative importance of neutral and adaptive processes (Gillespie 1994) to fundamental questions on levels of variation in natural populations (Lewontin and Hubby 1966), information harboured by proteins became an essential source of information. Even with advent of DNA sequencing, information carried at the protein level remains crucial, since most evolutionary tests rely either on estimates of rates of amino acid substitutions (Nei and Gojobori 1986; McDonald and Kreitman 1991) or their functional/structural effects (Woolley et al. 2003; Kelley et al. 2015). Thus, site-specific information from proteins (and underlying coding sequences) is a fundamental component of molecular evolution analyses, which can reveal signatures of the form and strength of selection.

Although we can infer a lot about the evolutionary history of individual proteins, they are not isolated entities. Proteins are embedded in pathways, which can constrain their 'freedom' to evolve (Fisher 1930) or create coevolutionary dynamics between interacting parts (Fraser et al. 2002). Moreover, the coding sequence of a protein can reflect the specific properties of the genome, rather than just the factors shaping the protein itself (D'Onofrio et al. 1991). For example, amino acids coded by AT-richer codons may appear more often than those coded by GC-richer codons in AT-biased genomes solely as a consequence of mutational bias (rather than as a consequence of protein function optimization). Furthermore, protein evolution can be constrained by selection to optimize usage of resources, measured in the currency of energetic costs (Akashi and Gojobori 2002). Therefore, selective pressures to reduce biosynthetic costs (which can limit usage of costly amino acids) can oppose selection to optimize protein function, resulting in proteins that reflect the outcome of this evolutionary compromise between function and cost (Swire 2007). Whereas recent studies have investigated the influence of protein networks on evolution of focal proteins (Fraser et al. 2002), the influence of broader processes shaping metabolic costs and genome composition are still poorly investigated.

Microbial eukaryotes provide powerful systems to investigate fundamental problems on molecular evolution. They share conserved pathways with more

complex eukaryotes, but with dimensionality reduced to one or a few cell subtypes (Bozzaro 2013). These organisms may have a strong potential for adaptive evolution, due to their large effective population sizes (Ohta 1992), at the same time that can have strongly biased base composition genomes, due to mutational bias (Sueoka 1988). The amoeba *Dictyostelium discoideum* is a model organism to understand mechanisms of cell signalling, motility, differentiation and de-differentiation, due to its simplicity and conservation of pathways across complex eukaryotes (Kessin 2001; Chisholm and Firtel 2004; Katoh et al. 2004; Bozzaro 2013). Its low complexity genome is characterized by very low GC content (~24%) and long stretches of amino acid repeats (Eichinger et al. 2005). This system also offers large scale gene expression data from various conditions (Parikh et al. 2010; Nasser et al. 2013; Rosengarten et al. 2015), as well as a collection of intra- and interspecific evolutionary parameters estimated from fully sequenced genomes (de Oliveira et al. in review). Thus, *D. discoideum* provides both an interesting biological system and a wide range of large-scale data.

At least three factors can, in principle, influence amino acid usage across the genome: number of codons, base composition, and metabolic costs. The number of synonymous codons may determine amino acid usage simply because more codons could result in a given amino acid appearing more frequently by random chance. Likewise, considering the base composition of the genome, some codons might be expected to be more common as a result of mutational biases. We tested this hypothesis by calculating the expected frequencies of codons under GC equilibrium (GC_{eq}), which provides an estimate of the neutral expectation under the assumption that codon frequencies are determined solely by mutational processes shaping base composition (for short, hereon referred simply as 'base composition'). Finally, costs of amino acid biosynthesis may impose constraints to amino acid usage, which could reduce usage of metabolically costly amino acids and favour usage of metabolically cheaper alternatives (Akashi and Gojobori 2002; Wagner 2005; Zhang et al. 2018). As many heterotrophs, however, *D. discoideum* obtains certain amino acids from their food and has lost the ability to synthesise 11 'essential' amino acids (namely, Arg, His, Ile, Leu, Lys, Met, Phe, Ser, Thr, Trp

and Val; (Payne and Loomis 2006)) – a feature that must be taken into account when analysing the influence metabolic costs on amino acid usage.

Using the reference genome of *D. discoideum* (Eichinger et al. 2005) to estimate amino acid frequencies, we fit linear models to understand factors influencing overall amino acid usage. A model considering these four factors plus an interaction between costs and the ability to synthesise an amino acid (model $M_{N+B+C+S+(C\times S)}$, Table 4.1) reveals that there is no significant effect of the number of codons ($P = 0.685$), or the ability of synthesise an amino acid ($P = 0.722$) or the interaction between this variable and metabolic cost ($P = 0.911$) on amino acid usage. Therefore, these predictors were excluded from further analysis. Interestingly, however, significance of metabolic costs with a lack of significance for both terms 'synthesis' and the interaction of this variable with costs suggests that the influence of metabolic cost in amino acid usage is generalized across both essential and non-essential amino acids.

A model including base composition and cost explains ~75% of overall amino acid usage (M_{B+C}, Table 4.1), increasing to ~87% with the addition of an interaction term between these two variables ($M_{B+C(B\times C)}$, Table 4.1). Although base composition alone explains a large variation of amino acid usage (M_B, Table 4.1), the model that includes both base composition and costs ($M_{B+C(B\times C)}$) fits significantly better than either M_B and a model that includes only costs (M_C, Table 4.1; ANOVA M_B versus $M_{B+C(B\times C)}$: $F = 27.17$, $P < 10^{-4}$; M_C versus $M_{B+C(B\times C)}$: $F = 49.36$, $P < 10^{-6}$). This model fits significantly for both groups of essential and non-essential amino acids, explaining 85% ($R^2 = 0.846$, $P = 0.007$) and 94% ($R^2 = 0.938$, $P = 0.002$) of variation in amino acid usage, respectively, supporting the hypothesis that cost is an important feature even for amino acids that are not synthesised, but obtained from food. This finding is consistent to general reports from heterotrophs (Swire 2007), where amino acid bioavailability constrains usage of metabolically expensive essential amino acids, because their synthesis is limited in the autotroph prey (bacteria, yeast) due to biosynthesis costs (Akashi and Gojobori 2002).

Table 4.1 Linear models explaining amino acid use across the genome.

Models with combinations of factors – number of alternative codons, base composition (expected frequencies under GC equilibrium (GC_{eq}), metabolic cost of amino acid biosynthesis, ability of synthesise the amino acid and interactions – were fitted to explain amino acid frequencies. Slope estimates of each variable included in the model are indicated, with significant values highlighted in bold.

Model	N	B	C	S	C×S	B×C	R^2	*P*
$M_{N+B+C+S+(C\times S)}$	−0.001	**0.531**	**−0.088**	−0.018	0.005	---	0.766	**< 10^{-3}**
M_{B+C}	---	**0.537**	**−0.073**			---	0.746	**< 10^{-4}**
$M_{B+C+(B\times C)}$	---	**2.408**	−0.033			**−1.214**	0.874	**< 10^{-6}**
M_B	---	**0.363**	---			---	0.415	**0.003**
M_C	---	---	−0.023			---	0.041	0.406

N: number of codons; B: base composition; C: metabolic costs; S: synthesised (Yes/No).

Because the *D. discoideum* genome is very AT-rich due to mutational bias (de Oliveira et al. in prep.b), these results suggest that amino acids coded by AT-richer codons are both predicted and observed to be used more often. Metabolic costs also explain a fraction of overall amino acid usage, but its effect could only be identified after accounting for the strong influence of base composition (compare models M_C and M_{B+C} in Table 1). Significance of an interaction term suggests that metabolic costs modulates the proportion of observed amino acid frequencies that can be explained by base composition. In fact, the influence of base composition on amino acid usage is maximum for metabolically 'cheaper' amino acids, but decreases with metabolic costs (Figure 4.1). Conversely, the relationship between observed frequencies and cost is stronger for amino acids predicted to be used very often, but decreases on those predicted to be rare from base composition (Figure 4.1). This pattern suggests that amino acids that are both metabolically cheaper and expected to be common from base composition (in this case, AT-richer codons) are in general used more often than costly and/or predicted to be rare from the underlying mutational process.

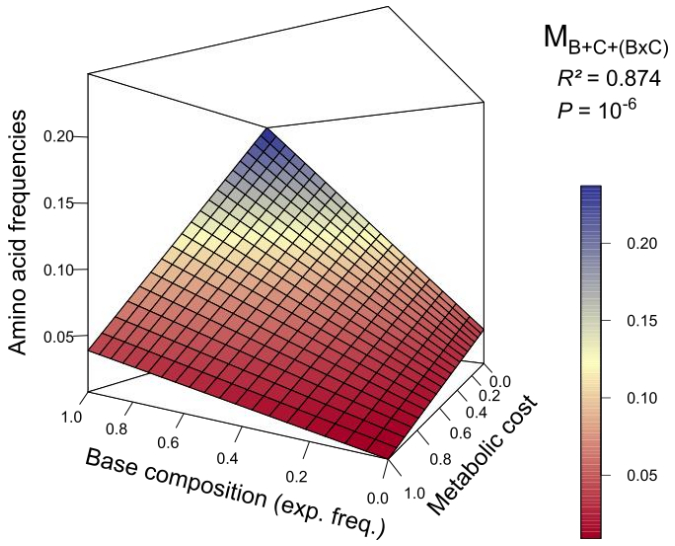

Figure 4.1 Visualisation of the influence of base composition and metabolic cost on the use of amino acids.

Overall amino acid usage can be mostly predicted by underlying processes shaping base composition (represented by the expected amino acid frequencies under GC_{eq}) and metabolic costs of amino acid biosynthesis ($\log_{10}$ scale, obtained from (Wagner 2005)). This landscape derives from fitted values of the model $M_{B+C+(B\times C)}$ (see Table 4.1), which explains ~87% of amino acid usage. The interaction term reflects that the relationship between processes shaping base composition and observed amino acid frequencies (colour scale) is modulated by metabolic cost.

Although base composition and metabolic costs explain a large proportion of the frequency with which different amino acids are used across the genome, their impact on individual proteins can be highly variable. Individual proteins vary widely in their specific properties and associated constraints, such as requiring specific amino acids for appropriate folding and function (Carugo 2008) or avoiding certain deleterious compositions, such as those that lead to formation of prion-like structures (Du 2011). To test to which extent amino acid composition of individual proteins conform to the overall pattern we see across the genome, we fitted the extended model of amino acid use predicted by base composition and cost (model $M_{B+C+(B\times C)}$ from Table 4.1) to each protein. Remarkably, this model fits significantly for 9764 proteins out of 12903, with a median R^2 = 0.598 (Figure S4.1).

It is possible that the proportion of amino acid composition predicted by base composition is driven by selection to match background processes shaping the nucleotide composition of the genome, but it is also possible that it emerges as a by-product of mutational biases. These two scenarios are expected to leave contrasting molecular evolutionary signatures. If amino acid use emerges passively as a by-product of mutational bias, we expect that genes where the use of amino acids more closely matches the pattern predicted by GC_{eq} should be under weaker selective constraint, and hence should show evolutionary signatures closer to the neutral expectation. Conversely, if patterns are driven by selection, we would expect these genes to reflect stronger evolutionary constraint, either with positive selection favouring mutations that result in amino acid use that conforms to the genome-wide pattern, or purifying selection removing variation that does not conform. To understand this problem, we categorized groups of genes with different levels of relative importance of base composition, cost, and the interaction term on the fit to the model $M_{B+C+(B\times C)}$, and estimated their median evolutionary rates and levels of polymorphism. Using information from 5509 orthologues, we find that all groups of genes evolve under purifying selection ($K_a/K_s < 1$), but both evolutionary rates (K_a/K_s) and rates of nonsynonymous substitutions (K_a) increase with stronger influence of base composition on amino acid usage ($R^2 = 0.779$, $P < 10^{-3}$ and $R^2 = 0.842$, $P < 10^{-3}$; Figures 4.2A and 4.2B). Moreover, intraspecific information from 12,809 coding sequences reveals that both the median nonsynonymous polymorphism and proportion of genes carrying at least one nonsense mutation are positively correlated to levels of relative importance of base composition on predicting amino acid composition ($R^2 = 0.797$, $P < 10^{-3}$ and $R^2 = 0.822$, $P < 10^{-3}$; Figures 4.2C and 4.2D). Conversely, signatures of stronger purifying selection increases with the relative importance of costs in explaining overall amino acid frequencies (Figures 4.2E-H and 4.2I-L), with the interaction term showing less clear patterns (Figure 4.2). These results suggest that the ability of our genome-level model to explain amino acid content in individual genes arises from a passive process, particularly those evolving under weak selection.

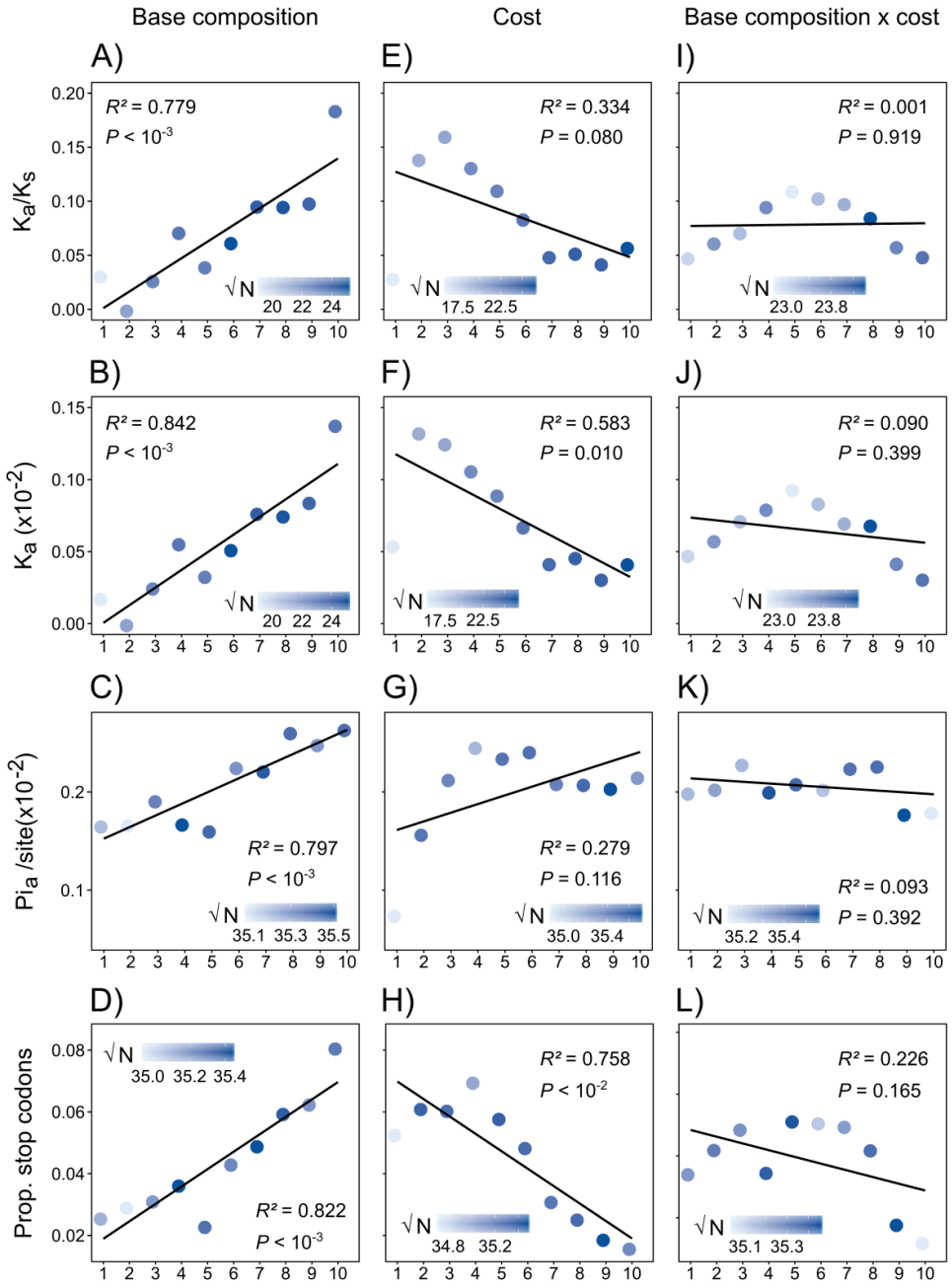

Figure 4.2 Evolutionary signatures of genes with different levels of relative importance of base composition, cost and the interaction factor on overall amino acid usage.

Proteins (Figure S4.1B) were divided in 10 bins (representing deciles) based on the relative importance of base composition (**A-D**), cost (**E-H**) and the interaction between these factors (**I-L**) on fit of the $M_{B+N+(B\times N)}$ model to individual proteins. Shown are the medians for evolutionary rates (**A**, **E** and **I**), rates of nonsynonymous substitutions (**B**, **F** and **J**), nonsynonymous average nucleotide diversity (**C**, **G** and **K**) and proportion of genes with at least one mutation introducing a premature stop codon (**D**, **H** and **L**). Lines represent fit to weighted regression models, where the weights are given by the squared root of the number of genes in the group.

A clear predicted consequence of the strong relationship between amino acid frequencies predicted based on base composition and costs with those observed in the genome would be a roughly homogeneous proteome in absence of selection on individual proteins. Expression is an important trait of individual genes/proteins, and has been implicated on variation in evolutionary rates – proteins required at high levels usually evolve more slowly and are strongly optimized to make usage of available resources (Duret and Mouchiroud 2000; Drummond et al. 2005; Akashi 2001). We therefore hypothesize that expression could be an important factor in determining amino acid composition of individual proteins because it increases the strength of selection on proteins relative to other genome-wide processes (such as those shaping base composition). To test this hypothesis, we used a dataset characterizing gene expression in vegetative and developmental stages of *D. discoideum* (Parikh et al. 2010; Nasser et al. 2013; Rosengarten et al. 2015), including expression levels of 11,918 genes (de Oliveira et al. in review). Expression is negatively correlated to the fraction of amino acid composition explained by genome-wide factors across proteins and explains ~32% of variation on this feature ($R^2 = 0.324$, $P < 10^{-15}$; Figure 4.3A). Because selection on expression optimization can encompass minimization of biosynthetic costs (Akashi and Gojobori 2002), we measured the relative importance of each variable of the model with increasing expression levels. Overall, the relative importance of base composition and the interaction between base composition and metabolic costs decreases across groups of genes with relatively higher expression levels, while the relative contribution of cost increases (Figure 4.3B) (becoming increasingly negative) (Figure S4.2). These results suggest that selection on expression features play an important role in shaping 'individuality' of amino acid composition in proteins by decreasing the effect of base composition, at the same time that optimization increases the overall (negative) effect of biosynthesis cost (i.e., the preferential use of lower cost amino acids).

To further test our hypothesis for the role of selection versus background processes, we analyse the relationship between influence of genome-wide factors shaping amino acid content and expression levels in two groups of genes evolving under different evolutionary constraints. Sociality genes evolve under Red King

dynamics characterized by weak selection due to conditional expression, whereas Non-sociality genes do not present this signature because they are expressed in every generation (de Oliveira et al. in review). In both groups of genes, the fraction of amino acid content explained by the genome-wide model (which is mostly affected by base composition) decrease with expression (Figure 4.3C), but this relationship is stronger in Sociality genes – exactly as predicted by our general findings.

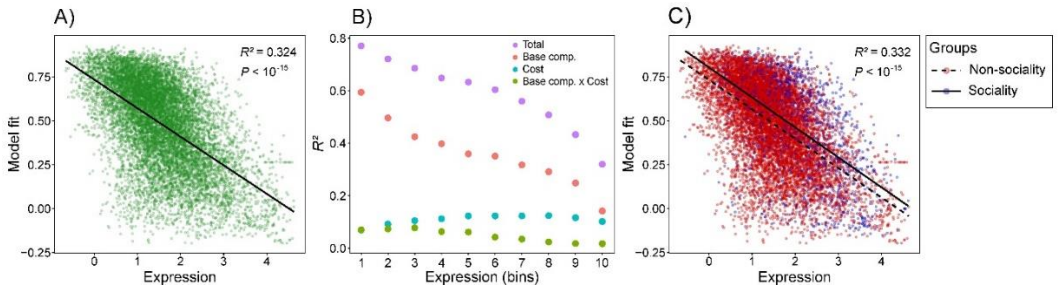

Figure 4.3 Expression and strong selection shapes individuality of protein amino acid content.

A) The fraction of amino acid content explained by genome-wide factors shaping genome base composition and metabolic costs decreases with expression levels. **B)** The relative contribution of cost and the interaction term between base composition and cost decreases with expression levels (divided in 10 groups representing deciles of the distribution of expression levels), whereas the relative importance of cost increases. **C)** Following expectations from our general results, fit to the genome-wide model decreases with expression levels, but it is still higher on genes evolving under diluted selection (Sociality genes) in comparison to genes that do not show this evolutionary signature (Non-sociality genes).

Our study reveals a striking influence of processes shaping genome and metabolic features on amino acid usage. This finding support previous discussions of the robustness of proteins (Kurland 1992), since the proteome remains functional even under the strong influence of these processes. A fine tune optimization, however, is essential for proteins required at high levels, consistent with a variety of studies showing slow evolution and optimization on highly expressed genes (Duret and Mouchiroud 2000; Drummond et al. 2005; Akashi 2001). Less intuitively, expression seems to be an essential feature that generates heterogeneity

in the proteome, since, in the absence of selection on individual proteins, the proteome would be essentially homogeneous. Taken together, these results highlight the importance of considering external effects shaping focal proteins, rather than viewing the protein as an isolated unity when studying molecular evolution.

4.2 Methods

4.2.1 Amino acid frequencies

Observed amino acid frequencies were estimated from reference genome (Eichinger et al. 2005) version 2.7 downloaded from Ensembl (Kersey et al. 2016), using seqinR (Charif and Lobry 2007). Because methionine is excluded from amino acid content analyses, frequencies were rescaled after excluding this amino acid.

4.2.2 Base composition and metabolic cost parameters

Influence of background processes was assessed by calculating expected amino acid frequencies were they solely a result of base composition, as described elsewhere (de Oliveira et al. in prep.b) and rescaled after removing methionine and stop codons. Metabolic costs of amino acid biosynthesis was obtained from (Wagner 2005). An approximate estimate of biosynthetic costs of each protein was obtained by calculation of the geometric mean for the protein sequence, defined as:

$$Cost_p = \left(\prod_{i=1}^{L} Cost_j \right)^{1/L}$$

where $Cost_j$ is the cost to synthesize amino acid j, and L is the protein sequence length.

4.2.3 Evolutionary tests

We used two data sets containing information from evolutionary rates and intraspecific evolutionary parameters (de Oliveira et al. in review). The first one includes rates of protein evolution (K_a/K_s) and rates of nonsynonymous substitutions (K_a) for 5509 orthologues estimated by comparison between the reference genome and a divergent *Dictyostelium* strain (OT3A). The second one contains several intraspecific parameters (such as π/site, number of nonsense mutations, etc.) for > 12,000 coding sequences, estimated from genome sequence data from 67 natural strains.

4.2.4 Gene expression

We used a dataset containing several gene expression parameters estimated from RNAseq data from vegetative (Parikh et al. 2010; Nasser et al. 2013; Rosengarten et al. 2015) and developmental stages (Rosengarten et al. 2015) to characterize the peak of maximum expression (in TMM units/sequence length) and whether the gene is conditional or expressed at every generation of vegetative growth (Sociality/Non-sociality genes, respectively) (de Oliveira et al. in review).

4.2.5 Statistical analyses

Statistical analyses were performed in R 3.5.1. Regressions and significance tests for the inclusion of each new regressor (ANOVA) were performed using base functions. Packages seqinR (Charif and Lobry 2007), ggplot2 (Wickham 2016) and relaimpo (Groemping 2006) were used to handling sequences, plotting and calculate relative importance of regressors in a model, respectively.

4.3 Supplementary material

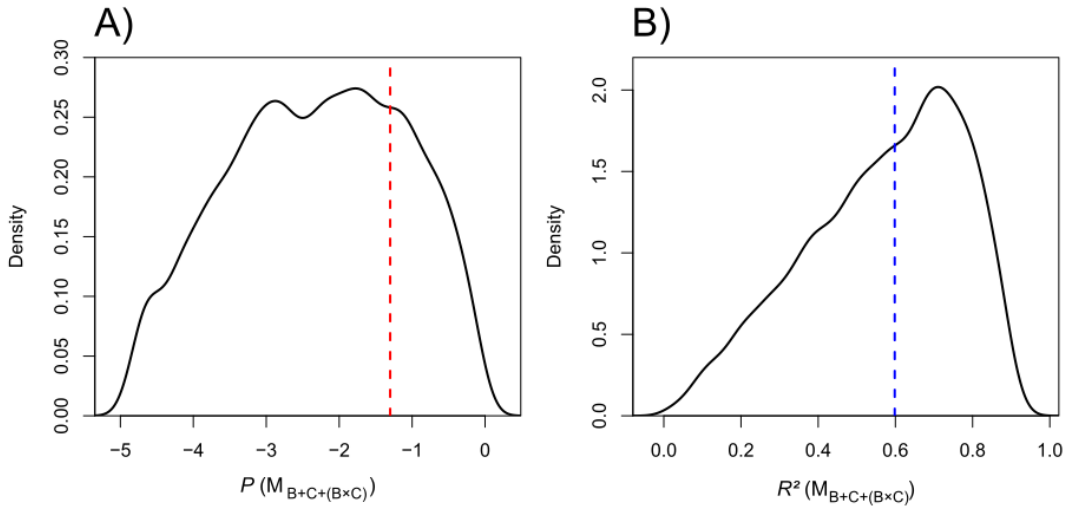

Figure S4.1 Fit of the model $M_{B+C+(B\times C)}$ of amino acid usage to proteins.

A model explaining amino acid content from base composition (expected frequencies under GC_{eq}) and metabolic costs was fitted to each protein. **A)** Distribution of *FDR*-corrected *P*-values of the model, with the dashed red line representing *FDR*-corrected P = 0.05. **B)** Distribution of model fit (R^2), with the dashed blue line representing the median value of $R^2 = 0.518$.

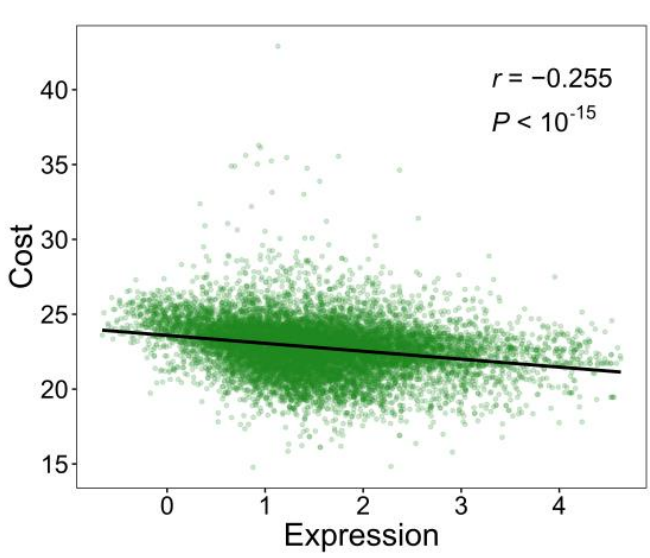

Figure S4.2 Correlation between biosynthesis costs and expression levels.

The geometric mean of metabolic costs to synthesize a protein is negatively correlated to expression levels.

5 General discussion

In this work I have investigated the relative contribution of natural selection and non-adaptive processes to the evolution of genes and the genome of a microbial eukaryote. Starting from the specific question about the evolutionary signatures harboured by genes underlying social traits, I contrasted four alternative scenarios for the evolution of 'social genes'. Although identified by different approaches, the four groups of social genes analysed show a unified signature of an evolutionary dynamic that we name the Red King process. This process is characterized by a dilution of the power of selection due to conditionality of the social cycle, which occurs only in a fraction of generations when the population starve. This finding is consistent to predictions from theoretical models (Linksvayer and Wade 2009; Van Dyken and Wade 2010, 2012) and with studies that investigate the evolutionary signatures of genes underlying social traits by contrasting adaptive and appropriate evolutionary null hypotheses (Van Dyken and Wade 2012; Warner et al. 2017).

Two groups of genes analysed in the second chapter were previously considered to evolve under conflict-driven dynamics (Ostrowski et al. 2015) or kin selection (Noh et al. 2018). However, conclusions were drawn from limited sets of genes, strains and analyses, sometimes including large genomic regions that may blur the signature carried by social genes, and more importantly, they lack a full consideration of evolutionary null hypotheses. For example, under the assumption that dynamics shaped by selection on social genes should affect the distribution of variation at linked sites, Ostrowski et al (2015) investigated signatures of selection in genomic windows containing mutations that disrupt cooperative behaviour (Santorelli et al. 2008). When large genomic windows are used (20Kb), patterns of linkage disequilibrium and high polymorphism are identified, but are diluted when window size is narrowed to half (10Kb). This suggests that the conclusion of balancing selection on these genes/mutations is unlikely to be associated to the target gene, particularly considering that recombination is high on this system (Flowers et al. 2010), and that in a 20Kb window size there are, on average, a total of four genes (Eichinger et al. 2005; Fey et al. 2013). When the analysis is restricted

to genes carrying such mutations, an excess of nonsynonymous polymorphism to divergence is identified in MK-tests, which was interpreted as a signature of balancing selection (Ostrowski et al. 2015). However, this test is influenced by segregation of slightly deleterious mutations (Parsch et al. 2009). When these two groups of social genes are included in our analyses (Figure 2.4), their evolutionary signatures are predicted simply by the proportion of conditionally expressed (sociality) genes within that group and are part of a larger scenario characterized by the RK process.

Findings from chapter 2 motivated further work in two different ways. First, it revealed that the dilution of selection caused by the RK process also affects synonymous codons, suggesting a role of selection in shaping synonymous codon usage. Second, it showed how overall features not directly related to selection in individual genes (in that case, conditional expression of the social cycle) can impact molecular evolution.

These two points were addresses in chapter 3, where I first characterized patterns of nucleotide substitution across the genome – identifying a strong bias towards AT accumulation – and analysed how this could impact the differential usage of alternative codons. Indeed, not only do all amino acids (and stop signal) use AT-rich codons much more often, but this parameter alone can explain a striking 88% variation of synonymous codon usage across the genome. This strong pattern could be inadvertently interpreted as a 'preference' towards usage of AT-richer codons. However, when contrasted against an appropriated null hypothesis of mutation-drift evolution, it reveals that the pattern is driven by non-adaptive mutational processes. After removing this effect, we identified sets of 'preferred' codons, whose use increases with expression and among genes evolving under stronger constraints, and are related to expression optimization by modulation of transcript stability.

Such a strong effect of mutation bias raised the hypothesis that this could have an impact on evolution at nonsynonymous sites as well and, consequently, amino acid usage. However, amino acid content can be also influenced by overall process shaping cell economics, such as minimization of costs of amino acid

biosynthesis (Akashi and Gojobori 2002; Wagner 2005). To understand how these processes shaping the genome and cell 'environments' can potentially affect amino acid content, we analysed amino acid usage across the proteome and in individual proteins. These two processes together explain a large fraction of amino acid frequencies in both proteome and protein levels. This is interesting because protein sequences are usually thought of as an intrinsic feature of a specific protein as a product of evolutionary processes acting at the protein level. Results presented in chapter 4 highlight the importance of considering the genome and cell contexts in which the protein is coded and expressed, rather than thinking the protein as an isolated entity.

The aim of this work is not, however, to argue against natural selection. For example, minimization of metabolic costs on amino acid biosynthesis is likely to be shaped by selection rather than non-adaptive processes – which is even clearer when analysed in the context of expression (Figures 4.3B and S4.2). Instead, I suggest a full consideration of evolutionary null hypothesis, especially when complex adaptive scenarios, such as social conflict, are also plausible.

Our work was designed and performed relying on hypothesis testing, i.e. contrasting alternative (adaptive) to null (neutral) scenarios, a core method in molecular evolutionary studies. Formulation of these alternative hypotheses was guided by current discussion in the literature, supported by evolutionary analyses (for example, in Chapter 2, where four alternative hypotheses were considered as putative evolutionary processes shaping social genes) and/or experimental work (for example, the role of certain codons in optimization of expression by modulation of transcript stability in Chapter 3). As a whole, our findings make several testable predictions about biological processes, which would certainly benefit from validation by follow-up experimentation.

6 References

Akashi, H. 1995. “Inferring Weak Selection from Patterns of Polymorphism and Divergence at ‘Silent’ Sites in *Drosophila* DNA.” *Genetics* 139 (2): 1067–76.

———. 2001. “Gene Expression and Molecular Evolution.” *Current Opinion in Genetics & Development* 11 (6): 660–66. https://doi.org/10.1016/S0959-437X(00)00250-1.

Akashi, H., and A. Eyre-Walker. 1998. “Translational Selection and Molecular Evolution.” *Current Opinion in Genetics & Development* 8 (6): 688–93. https://doi.org/10.1016/S0959-437X(98)80038-5.

Akashi, H., and T. Gojobori. 2002. “Metabolic Efficiency and Amino Acid Composition in the Proteomes of *Escherichia coli* and *Bacillus subtilis*.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (6): 3695–3700. https://doi.org/10.1073/pnas.062526999.

Akashi, H., N. Osada, and T. Ohta. 2012. “Weak Selection and Protein Evolution.” *Genetics* 192 (1): 15–31. https://doi.org/10.1534/genetics.112.140178.

Aken, B. L., S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. F. Banet, et al. 2016. “The Ensembl Gene Annotation System.” *Database* 2016 (January). https://doi.org/10.1093/database/baw093.

Aquadro, C. F., K. M. Lado, and W. A. Noon. 1988. “The Rosy Region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting Levels of Naturally Occurring DNA Restriction Map Variation and Divergence.” *Genetics* 119 (4): 875–88.

Barker, M. S., J. P. Demuth, and M. J. Wade. 2005. “Maternal Expression Relaxes Constraint on Innovation of the Anterior Determinant, *Bicoid*.” *PLOS Genetics* 1 (5): e57. https://doi.org/10.1371/journal.pgen.0010057.

Benabentos, R., S. Hirose, R. Sucgang, T. Curk, M. Katoh, E. A. Ostrowski, J. E. Strassmann, et al. 2009. “Polymorphic Members of the *Lag* Gene Family Mediate Kin Discrimination in *Dictyostelium*.” *Current Biology: CB* 19 (7): 567–72. https://doi.org/10.1016/j.cub.2009.02.037.

Berleth, T., M. Burri, G. Thoma, D. Bopp, S. Richstein, G. Frigerio, M. Noll, and C. Nüsslein-Volhard. 1988. “The Role of Localization of *Bicoid* RNA in Organizing the Anterior Pattern of the *Drosophila* Embryo.” *The EMBO Journal* 7 (6): 1749–56.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. “The Mosaic Genome of Warm-Blooded Vertebrates.” *Science* 228 (4702): 953–58. https://doi.org/10.1126/science.4001930.

Bernardi, G. 2000. “Isochores and the Evolutionary Genomics of Vertebrates.” *Gene* 241 (1): 3–17. https://doi.org/10.1016/S0378-1119(99)00485-0.

Bijma, P., W. M. Muir, E. D. Ellen, J. B. Wolf, and J. A. M. Van Arendonk. 2007. “Multilevel Selection 2: Estimating the Genetic Parameters Determining Inheritance and Response to Selection.” *Genetics* 175 (1): 289–99. https://doi.org/10.1534/genetics.106.062729.Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for

Illumina Sequence Data." *Bioinformatics (Oxford, England)* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bozzaro, Salvatore. 2013. "The Model Organism <Emphasis Type="Italic">Dictyostelium Discoideum</Emphasis>." In *Dictyostelium Discoideum Protocols*, 17–37. Methods in Molecular Biology. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-302-2_2.

Bozzaro, S. 2013. "The Model Organism *Dictyostelium discoideum*." In *Dictyostelium Discoideum Protocols*, 17–37. Methods in Molecular Biology. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-302-2_2.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Brisson, J. A., and S. V. Nuzhdin. 2008. "Rarity of Males in Pea Aphids Results in Mutational Decay." *Science (New York, N.Y.)* 319 (5859): 58. https://doi.org/10.1126/science.1147919.

Brockhurst, M. A., T. Chapman, K. C. King, J. E. Mank, S. Paterson, and G. D. D. Hurst. 2014. "Running with the Red Queen: The Role of Biotic Conflicts in Evolution." *Proceedings of the Royal Society of London B: Biological Sciences* 281 (1797): 20141382. https://doi.org/10.1098/rspb.2014.1382.

Bushnell, B. 2016. "BBMap Short-Read Aligner, and Other Bioinformatics Tools." 2016. https://sourceforge.net/projects/bbmap/.

Buttery, N. J., C. R. L. Thompson, and J. B. Wolf. 2010. "Complex Genotype Interactions Influence Social Fitness during the Developmental Phase of the Social Amoeba *Dictyostelium discoideum*." *Journal of Evolutionary Biology* 23 (8): 1664–71. https://doi.org/10.1111/j.1420-9101.2010.02032.x.

Buttery, N. J., D. E. Rozen, J. B. Wolf, and C. R. L. Thompson. 2009. "Quantification of Social Behavior in *D. Discoideum* Reveals Complex Fixed and Facultative Strategies." *Current Biology* 19 (16): 1373–77. https://doi.org/10.1016/j.cub.2009.06.058.

Carugo, O. 2008. "Amino Acid Composition and Protein Dimension." *Protein Science : A Publication of the Protein Society* 17 (12): 2187–91. https://doi.org/10.1110/ps.037762.108.

Castillo, D. I., G. T. Switz, K. R. Foster, D. C. Queller, and J. E. Strassmann. 2005. "A Cost to Chimerism in *Dictyostelium discoideum* on Natural Substrates." *Evolutionary Ecology Research* 7 (2): 263–71.

Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. "Hearing Silence: Non-Neutral Evolution at Synonymous Sites in Mammals." *Nature Reviews Genetics* 7 (2): 98–108. https://doi.org/10.1038/nrg1770.

Chamary, J. V., and L. D. Hurst. 2005. "Evidence for Selection on Synonymous Mutations Affecting Stability of mRNA Secondary Structure in Mammals." *Genome Biology* 6 (9): R75. https://doi.org/10.1186/gb-2005-6-9-r75.

Charif, D., and J. R. Lobry. 2007. "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis." In *Structural Approaches to Sequence Evolution*, edited by Dr Ugo Bastolla, Professor Dr Markus Porto, Dr H. Eduardo Roman, and Dr Michele Vendruscolo, 207–32. Biological and Medical

Physics, Biomedical Engineering. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35306-5_10.

Charlesworth, B., J. A. Coyne, and N. H. Barton. 1987. “The Relative Rates of Evolution of Sex Chromosomes and Autosomes.” *The American Naturalist* 130 (1): 113–46.

Chattwood, A., K. Nagayama, P. Bolourani, L. Harkin, M. Kamjoo, G. Weeks, and C. R. L. Thompson. 2013. “Developmental Lineage Priming in *Dictyostelium* by Heterogeneous *Ras* Activation.” *ELife* 2. https://doi.org/10.7554/eLife.01067.

Chen, S. L., W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. 2004. “Codon Usage between Genomes Is Constrained by Genome-Wide Mutational Processes.” *Proceedings of the National Academy of Sciences* 101 (10): 3480–85. https://doi.org/10.1073/pnas.0307827100.

Chisholm, R. L., and R. A. Firtel. 2004. “Insights into Morphogenesis from a Simple Developmental System.” *Nature Reviews Molecular Cell Biology* 5 (7): 531–41. https://doi.org/10.1038/nrm1427.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff.” *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Crick, F. H. C., L. Barnett, S. Brenner, and R. J. Watts-Tobin. 1961. “General Nature of the Genetic Code for Proteins.” *Nature* 192 (4809): 1227–32. https://doi.org/10.1038/1921227a0.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics (Oxford, England)* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Demuth, J. P., and M. J. Wade. 2007. “Maternal Expression Increases the Rate of *Bicoid* Evolution by Relaxing Selective Constraint.” *Genetica* 129 (1): 37–43. https://doi.org/10.1007/s10709-006-0031-4.

de Oliveira, J. L., A. C. Morales, B. Stewart, N. Gruenheit, S. B. Brow, R. A. de Brito, L. D. Hurst, A. O. Urrutia, C. R. L. Thompson, and J. B. Wolf. In review. “The Red King Process Explains Molecular Evolution of Social Genes in a Microbe.”

de Oliveira, J. L., A. C. Morales, L. D. Hurst, A. O. Urrutia, C. R. L. Thompson, and J. B. Wolf. In prep.a. “Amino Acid Composition Is Influenced by Evolutionary Processes Shaping Genomic and Metabolic Features in a Microbe.”

de Oliveira, J. L., A. C. Morales, L. D. Hurst, A. O. Urrutia, C. R. L. Thompson, and J. B. Wolf. In prep.b “Processes Shaping Synonymous Codon Use in an Extremely AT-Biased Genome.”

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5): 491–98. https://doi.org/10.1038/ng.806.

D’Onofrio, G., D. Mouchiroud, B. Aïssani, C. Gautier, and G. Bernardi. 1991. “Correlations between the Compositional Properties of Human Genes,

Codon Usage, and Amino Acid Composition of Proteins." *Journal of Molecular Evolution* 32 (6): 504–10. https://doi.org/10.1007/BF02102652.

dos Reis, M., R. Savva, and L. Wernisch. 2004. "Solving the Riddle of Codon Usage Preferences: A Test for Translational Selection." *Nucleic Acids Research* 32 (17): 5036–44. https://doi.org/10.1093/nar/gkh834.

dos Reis, M., L. Wernisch, and R. Savva. 2003. "Unexpected Correlations between Gene Expression and Codon Usage Bias from Microarray Data for the Whole *Escherichia coli* K-12 Genome." *Nucleic Acids Research* 31 (23): 6976–85. https://doi.org/10.1093/nar/gkg897.

dos Santos, G., A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, and W. M. Gelbart. 2015. "FlyBase: Introduction of the *Drosophila melanogaster* Release 6 Reference Genome Assembly and Large-Scale Migration of Genome Annotations." *Nucleic Acids Research* 43 (Database issue): D690–97. https://doi.org/10.1093/nar/gku1099.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. "Why Highly Expressed Proteins Evolve Slowly." *Proceedings of the National Academy of Sciences of the United States of America* 102 (40): 14338–43. https://doi.org/10.1073/pnas.0504070102.

Drummond, D. A., and C. O. Wilke. 2009. "The Evolutionary Consequences of Erroneous Protein Synthesis." *Nature Reviews. Genetics* 10 (10): 715–24. https://doi.org/10.1038/nrg2662.

Du, Z. 2011. "The Complexity and Implications of Yeast Prion Domains." *Prion* 5 (4): 311–16. https://doi.org/10.4161/pri.5.4.18304.

Dugatkin, L. A.. 1998. "Game Theory and Cooperation." In *Game Theory and Animal Behavior*, New Ed edition, 336. New York Oxford: Oxford University Press USA.

Duret, L., and D. Mouchiroud. 2000. "Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate." *Molecular Biology and Evolution* 17 (1): 68–070. https://doi.org/10.1093/oxfordjournals.molbev.a026239.

Eichinger, L., J. A. Pachebat, G. Glöckner, M.-A. Rajandream, R. Sucgang, M. Berriman, J. Song, et al. 2005. "The Genome of the Social Amoeba *Dictyostelium discoideum*." *Nature* 435 (7038): 43–57. https://doi.org/10.1038/nature03481.

Ellegren, H. 2004. "Microsatellites: Simple Sequences with Complex Evolution." *Nature Reviews Genetics* 5 (6): 435–45. https://doi.org/10.1038/nrg1348.

Ellegren, H., and N. Galtier. 2016. "Determinants of Genetic Diversity." *Nature Reviews Genetics* 17 (7): 422–33. https://doi.org/10.1038/nrg.2016.58.

Fey, P., R. J. Dodson, S. Basu, and R. L. Chisholm. 2013. "One Stop Shop for Everything *Dictyostelium*: DictyBase and the Dicty Stock Center in 2012." *Methods in Molecular Biology (Clifton, N.J.)* 983: 59–92. https://doi.org/10.1007/978-1-62703-302-2_4.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press. https://www.biodiversitylibrary.org/bibliography/27468.

Flowers, J. M., Si I. L., A. Stathos, G. Saxer, E. A. Ostrowski, D. C. Queller, J. E. Strassmann, and M. D. Purugganan. 2010. "Variation, Sex, and Social Cooperation: Molecular Population Genetics of the Social Amoeba

Dictyostelium discoideum." *PLOS Genet* 6 (7): e1001013. https://doi.org/10.1371/journal.pgen.1001013.

Fortunato, A., J. E. Strassmann, L. Santorelli, and D. C. Queller. 2003. "Co-Occurrence in Nature of Different Clones of the Social Amoeba, *Dictyostelium discoideum*." *Molecular Ecology* 12 (4): 1031–38. https://doi.org/10.1046/j.1365-294X.2003.01792.x.

Foster, K. R. 2006. "Sociobiology: The Phoenix Effect." *Nature* 441 (7091): 291. https://doi.org/10.1038/441291a.

———. 2010. "Social Behaviour in Microorganisms." In *Social Behaviour: Genes, Ecology and Evolution*, 1st edition, 574. Cambridge ; New York: Cambridge University Press.

Foster, K. R., A. Fortunato, J. E. Strassmann, and D. C. Queller. 2002. "The Costs and Benefits of Being a Chimera." *Proceedings of the Royal Society of London B: Biological Sciences* 269 (1507): 2357–62. https://doi.org/10.1098/rspb.2002.2163.

Foster, K. R., K. Parkinson, and C. R. L. Thompson. 2007. "What Can Microbial Genetics Teach Sociobiology?" *Trends in Genetics: TIG* 23 (2): 74–80. https://doi.org/10.1016/j.tig.2006.12.003.

Foster, K. R., G. Shaulsky, J. E. Strassmann, D. C. Queller, and C. R. L. Thompson. 2004. "Pleiotropy as a Mechanism to Stabilize Cooperation." *Nature* 431 (7009): 693–96. https://doi.org/10.1038/nature02894.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. "Evolutionary Rate in the Protein Interaction Network." *Science* 296 (5568): 750–52. https://doi.org/10.1126/science.1068696.

Fraser, Hunter B., Aaron E. Hirsh, Lars M. Steinmetz, Curt Scharfe, and Marcus W. Feldman. 2002. "Evolutionary Rate in the Protein Interaction Network." *Science* 296 (5568): 750–52. https://doi.org/10.1126/science.1068696.

Galtier, N. 2016. "Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis." *PLOS Genet* 12 (1): e1005774. https://doi.org/10.1371/journal.pgen.1005774.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, et al. 2002. "Genome Sequence of the Human Malaria Parasite *Plasmodium falciparum*." *Nature* 419 (6906): 498–511. https://doi.org/10.1038/nature01097.

Garneau, N. L., J. W., and C. J. Wilusz. 2007. "The Highways and Byways of mRNA Decay." *Nature Reviews. Molecular Cell Biology* 8 (2): 113–26. https://doi.org/10.1038/nrm2104.

Gilbert, O. M., K. R. Foster, N. J. Mehdiabadi, J. E. Strassmann, and D. C. Queller. 2007. "High Relatedness Maintains Multicellular Cooperation in a Social Amoeba by Controlling Cheater Mutants." *Proceedings of the National Academy of Sciences of the United States of America* 104 (21): 8913–17. https://doi.org/10.1073/pnas.0702723104.

Gilbert, O. M., D. C. Queller, and J. E. Strassmann. 2009. "Discovery of a Large Clonal Patch of a Social Amoeba: Implications for Social Evolution." *Molecular Ecology* 18 (6): 1273–81. https://doi.org/10.1111/j.1365-294X.2009.04108.x.

Gillespie, J. H. 1994. *The Causes of Molecular Evolution*. New Ed edition. New York: Oxford University Press.

Gingold, H., and Y. Pilpel. 2011. “Determinants of Translation Efficiency and Accuracy.” *Molecular Systems Biology* 7 (April): 481. https://doi.org/10.1038/msb.2011.14.

Glöckner, G., H. M. Lawal, M. Felder, R. Singh, G. Singer, C. J. Weijer, and P. Schaap. 2016. “The Multicellularity Genes of Dictyostelid Social Amoebas.” *Nature Communications* 7 (June): 12085. https://doi.org/10.1038/ncomms12085.

Grantham, R. 1980. “Working of the Genetic Code.” *Trends in Biochemical Sciences* 5 (12): 327–31. https://doi.org/10.1016/0968-0004(80)90143-7.

Groemping, U. 2006. “Relative Importance for Linear Regression in R: The Package Relaimpo.” *Journal of Statistical Software* 17. http://dx.doi.org/10.18637/jss.v017.i01.

Gruenheit, N., K. Parkinson, B. Stewart, J. A. Howie, J. B. Wolf, and C. R. L. Thompson. 2017. “A Polychromatic ‘Greenbeard’ Locus Determines Patterns of Cooperation in a Social Amoeba.” *Nature Communications* 8: 14171. https://doi.org/10.1038/ncomms14171.

Haldane, J. B. S. 1957. “The Cost of Natural Selection.” *Journal of Genetics* 55 (3): 511. https://doi.org/10.1007/BF02984069.

Hamilton, W. D. 1964a. “The Genetical Evolution of Social Behaviour. I.” *Journal of Theoretical Biology* 7 (1): 1–16.

———. 1964b. “The Genetical Evolution of Social Behaviour. II.” *Journal of Theoretical Biology* 7 (1): 17–52.

Harris, W. E., A. J. McKane, and J. B. Wolf. 2008. “The Maintenance of Heritable Variation through Social Competition.” *Evolution; International Journal of Organic Evolution* 62 (2): 337–47. https://doi.org/10.1111/j.1558-5646.2007.00302.x.

Hughes, A. L. 2008. “Near-Neutrality: The Leading Edge of the Neutral Theory of Molecular Evolution.” *Annals of the New York Academy of Sciences* 1133: 162–79. https://doi.org/10.1196/annals.1438.001.

Hurst, L. D. 2002. “The Ka/Ks Ratio: Diagnosing the Form of Sequence Evolution.” *Trends in Genetics* 18 (9): 486–87. https://doi.org/10.1016/S0168-9525(02)02722-1.

Husemann, M., F. E. Zachos, R. J. Paxton, and J. C. Habel. 2016. “Effective Population Size in Ecology and Evolution.” *Heredity* 117 (4): 191–92. https://doi.org/10.1038/hdy.2016.75.

Ikemura, T. 1985. “Codon Usage and tRNA Content in Unicellular and Multicellular Organisms.” *Molecular Biology and Evolution* 2 (1): 13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335.

Ikemura, T. 1981. “Correlation between the Abundance of *Escherichia coli* Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes: A Proposal for a Synonymous Codon Choice That Is Optimal for the *E. coli* Translational System.” *Journal of Molecular Biology* 151 (3): 389–409. https://doi.org/10.1016/0022-2836(81)90003-6.

Jack, C. N., N. Buttery, B. Adu-Oppong, M. Powers, C. R. L. Thompson, D. C. Queller, and J. E. Strassmann. 2015. “Migration in the Social Stage of *Dictyostelium discoideum* Amoebae Impacts Competition.” *PeerJ* 3: e1352. https://doi.org/10.7717/peerj.1352.

Jiang, H., R. Lei, S.-W. Ding, and S. Zhu. 2014. “Skewer: A Fast and Accurate Adapter Trimmer for next-Generation Sequencing Paired-End Reads.” *BMC Bioinformatics* 15: 182. https://doi.org/10.1186/1471-2105-15-182.

Johnson, N. A., and J. Lachance. 2012. “The Genetics of Sex Chromosomes: Evolution and Implications for Hybrid Incompatibility.” *Annals of the New York Academy of Sciences* 1256: E1-22. https://doi.org/10.1111/j.1749-6632.2012.06748.x.

Kapheim, K. M., H. Pan, C. Li, S. L. Salzberg, D. Puiu, T. Magoc, H. M. Robertson, et al. 2015. “Genomic Signatures of Evolutionary Transitions from Solitary to Group Living.” *Science* 348 (6239): 1139–43. https://doi.org/10.1126/science.aaa4788.

Katoh, M., C. Shaw, Q. Xu, N. Van Driessche, T. Morio, H. Kuwayama, S. Obara, H. Urushihara, Y. Tanaka, and G. Shaulsky. 2004. “An Orderly Retreat: Dedifferentiation Is a Regulated Process.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (18): 7005–10. https://doi.org/10.1073/pnas.0306983101.

Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass, and M. J. E. Sternberg. 2015. “The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis.” *Nature Protocols* 10 (6): 845–58. https://doi.org/10.1038/nprot.2015.053.

Kersey, P. J., J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, et al. 2016. “Ensembl Genomes 2016: More Genomes, More Complexity.” *Nucleic Acids Research* 44 (D1): D574-580. https://doi.org/10.1093/nar/gkv1209.

Kessin, R. H. 2001. *Dictyostelium: Evolution, Cell Biology, and the Development of Multicellularity*. Cambridge University Press.

Kimura, M. 1968. “Genetic Variability Maintained in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Isoalleles.” *Genetical Research* 11 (3): 247–69.

Kimura, M. 1968. “Genetic Variability Maintained in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Isoalleles.” *Genetical Research* 11 (3): 247–69.

Kimura, M. 1968. “Evolutionary Rate at the Molecular Level.” *Nature* 217 (5129): 624–26. https://doi.org/10.1038/217624a0.

———. 1969. “The Rate of Molecular Evolution Considered from the Standpoint of Population Genetics.” *Proceedings of the National Academy of Sciences of the United States of America* 63 (4): 1181–88.

———. 1985. *The Neutral Theory of Molecular Evolution*. Revised ed. edition. Cambridge Cambridgeshire; New York: Cambridge University Press.

Kimura, M., and T. Ohta. 1971. “Protein Polymorphism as a Phase of Molecular Evolution.” *Nature* 229 (5285): 467–69. https://doi.org/10.1038/229467a0.

King, J. L., and T. H. Jukes. 1969. “Non-Darwinian Evolution.” *Science* 164 (3881): 788–98. https://doi.org/10.1126/science.164.3881.788.

Kirkup, B. C., and M. A. Riley. 2004. “Antibiotic-Mediated Antagonism Leads to a Bacterial Game of Rock–Paper–Scissors *in Vivo*.” *Nature* 428 (6981): 412. https://doi.org/10.1038/nature02429.

Kudla, G., L. Lipinski, F. Caffin, A. Helwak, and M. Zylicz. 2006. “High Guanine and Cytosine Content Increases MRNA Levels in Mammalian Cells.” *PLoS Biology* 4 (6). https://doi.org/10.1371/journal.pbio.0040180.

Kurland, C. G. 1992. “Translational Accuracy and the Fitness of Bacteria.” *Annual Review of Genetics* 26: 29–50. https://doi.org/10.1146/annurev.ge.26.120192.000333.

Laird, C. D., B. L. McConaughy, and B. J. McCarthy. 1969. “Rate of Fixation of Nucleotide Substitutions in Evolution.” *Nature* 224 (5215): 149–54.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, and M. Przeworski. 2012. “Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species?” *PLoS Biology* 10 (9): e1001388. https://doi.org/10.1371/journal.pbio.1001388.

Lewontin, R. C. 1974. *The Genetic Basis of Evolutionary Change*. New York: Columbia Univerisity Press.

Lewontin, R. C., and J. L. Hubby. 1966. “A Molecular Approach to the Study of Genic Heterozygosity in Natural Populations. II. Amount of Variation and Degree of Heterozygosity in Natural Populations of *Drosophila pseudoobscura*.” *Genetics* 54 (2): 595–609.

Li, S. I., N. J. Buttery, C. R. L. Thompson, and M. D. Purugganan. 2014. “Sociogenomics of Self vs. Non-Self Cooperation during Development of *Dictyostelium discoideum*.” *BMC Genomics* 15 (July): 616. https://doi.org/10.1186/1471-2164-15-616.

Li, S. I., and M. D. Purugganan. 2011. “The Cooperative Amoeba: *Dictyostelium* as a Model for Social Evolution.” *Trends in Genetics: TIG* 27 (2): 48–54. https://doi.org/10.1016/j.tig.2010.11.003. Linksvayer, T. A., and M. J. Wade. 2009. “Genes with Social Effects Are Expected to Harbor More Sequence Variation within and between Species.” *Evolution; International Journal of Organic Evolution* 63 (7): 1685–96. https://doi.org/10.1111/j.1558-5646.2009.00670.x.

———. 2016. “Theoretical Predictions for Sociogenomic Data: The Effects of Kin Selection and Sex-Limited Expression on the Evolution of Social Insect Genomes.” *Frontiers in Ecology and Evolution* 4. https://doi.org/10.3389/fevo.2016.00065.

Lorenz, R., S. H. Bernhart, C. H. zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. 2011. “ViennaRNA Package 2.0.” *Algorithms for Molecular Biology* 6 (1): 26. https://doi.org/10.1186/1748-7188-6-26.

Lu, J., and C.-I. Wu. 2005. “Weak Selection Revealed by the Whole-Genome Comparison of the X Chromosome and Autosomes of Human and Chimpanzee.” *Proceedings of the National Academy of Sciences* 102 (11): 4063–67. https://doi.org/10.1073/pnas.0500436102.

Lubin, Y., and T. Bilde. 2007. “The Evolution of Sociality in Spiders.” In *Advances in the Study of Behavior*, 37:83–145. Academic Press. https://doi.org/10.1016/S0065-3454(07)37003-4.

Madgwick, P. G., B. Stewart, L. J. Belcher, C. R. L. Thompson, and J. B. Wolf. 2018. “Strategic Investment Explains Patterns of Cooperation and Cheating in a Microbe.” *Proceedings of the National Academy of Sciences* 115 (21): E4823–32. https://doi.org/10.1073/pnas.1716087115.

Mank, J. E., E. Axelsson, and H. Ellegren. 2007. “Fast-X on the Z: Rapid Evolution of Sex-Linked Genes in Birds.” *Genome Research* 17 (5): 618–24. https://doi.org/10.1101/gr.6031907.

Maynard-Smith, J., and G. R. Price. 1973. “The Logic of Animal Conflict.” *Nature* 246 (5427): 15–18. https://doi.org/10.1038/246015a0.

McDonald, J. H., and M. Kreitman. 1991. “Adaptive Protein Evolution at the *Adh* Locus in *Drosophila*.” *Nature* 351 (6328): 652–54. https://doi.org/10.1038/351652a0.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, et al. 2007. “The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions.” *Science (New York, N.Y.)* 318 (5848): 245–50. https://doi.org/10.1126/science.1143609.

Miller, M. B., and B. L. Bassler. 2001. “Quorum Sensing in Bacteria.” *Annual Review of Microbiology* 55 (1): 165–99. https://doi.org/10.1146/annurev.micro.55.1.165.

Moore, A. J., E. D. Brodie III, and J. B. Wolf. 1997. “Interacting Phenotypes and the Evolutionary Process: I. Direct and Indirect Genetic Effects of Social Interactions.” *Evolution* 51 (5): 1352–62. https://doi.org/10.2307/2411187.

Moqtaderi, Z., J. V. Geisberg, and K. Struhl. 2014. “Secondary Structures Involving the Poly(A) Tail and Other 3’ Sequences Are Major Determinants of mRNA Isoform Stability in Yeast.” *Microbial Cell* 1 (4): 137–39. https://doi.org/10.15698/mic2014.04.140.

Muñoz-Dorado, J., and J. M. Arias. 1995. “The Social Behavior of Myxobacteria.” *Microbiologia (Madrid, Spain)* 11 (4): 429–38.

Muto, A., and S. Osawa. 1987. “The Guanine and Cytosine Content of Genomic DNA and Bacterial Evolution.” *Proceedings of the National Academy of Sciences* 84 (1): 166–69. https://doi.org/10.1073/pnas.84.1.166.

Nakabachi, A., A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M. Hattori. 2006. “The 160-Kilobase Genome of the Bacterial Endosymbiont *Carsonella*.” *Science* 314 (5797): 267–267. https://doi.org/10.1126/science.1134196.

Nasser, W., B. Santhanam, E. R. Miranda, A. Parikh, K. Juneja, G. Rot, C. Dinh, et al. 2013. “Bacterial Discrimination by Dictyostelid Amoebae Reveals the Complexity of Ancient Interspecies Interactions.” *Current Biology : CB* 23 (10): 862–72. https://doi.org/10.1016/j.cub.2013.04.034.

Nei, M., and T. Gojobori. 1986. “Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions.” *Molecular Biology and Evolution* 3 (5): 418–26.

Nei, M. 2005. “Selectionism and Neutralism in Molecular Evolution.” *Molecular Biology and Evolution* 22 (12): 2318–42. https://doi.org/10.1093/molbev/msi242.

———. 2013. *Mutation-Driven Evolution*. 1st ed. Oxford, United Kingdom: Oxford University Press.

Nei, M., Y. Suzuki, and M. Nozawa. 2010. “The Neutral Theory of Molecular Evolution in the Genomic Era.” *Annual Review of Genomics and Human Genetics* 11 (1): 265–89. https://doi.org/10.1146/annurev-genom-082908-150129.

Nirenberg, M., T. Caskey, R. Marshall, R. Brimacombe, D. Kellogg, B. Doctor, D. Hatfield, et al. 1966. “The RNA Code and Protein Synthesis.” *Cold Spring Harbor Symposia on Quantitative Biology* 31 (January): 11–24. https://doi.org/10.1101/SQB.1966.031.01.008.

Noh, S., K. S. Geist, X. Tian, J. E. Strassmann, and D. C. Queller. 2018. “Genetic Signatures of Microbial Altruism and Cheating in Social Amoebas in the Wild.” *Proceedings of the National Academy of Sciences*, March, 201720324. https://doi.org/10.1073/pnas.1720324115.

Ohta, T. 1973. “Slightly Deleterious Mutant Substitutions in Evolution.” *Nature* 246 (5428): 96–98. https://doi.org/10.1038/246096a0.

———. 1992. “The Nearly Neutral Theory of Molecular Evolution.” *Annual Review of Ecology and Systematics* 23 (1): 263–86. https://doi.org/10.1146/annurev.es.23.110192.001403.

———. 2012. “Tomoko Ohta.” *Current Biology* 22 (16): R618–19. https://doi.org/10.1016/j.cub.2012.06.031.

Ohta, T., and J. H. Gillespie. 1996. “Development of Neutral and Nearly Neutral Theories.” *Theoretical Population Biology* 49 (2): 128–42. https://doi.org/10.1006/tpbi.1996.0007.

Ostrowski, E. A., M. Katoh, G. Shaulsky, D. C. Queller, and J. E. Strassmann. 2008. “Kin Discrimination Increases with Genetic Distance in a Social Amoeba.” *PLOS Biology* 6 (11): e287. https://doi.org/10.1371/journal.pbio.0060287.

Ostrowski, E. A., Y. Shen, X. Tian, R. Sucgang, H. Jiang, J. Qu, M. Katoh-Kurasawa, et al. 2015. “Genomic Signatures of Cooperation and Conflict in the Social Amoeba.” *Current Biology: CB* 25 (12): 1661–65. https://doi.org/10.1016/j.cub.2015.04.059.

Page, R. D. M., and E. C. Holmes. 1998. *Molecular Evolution: A Phylogenetic Approach*. 1st ed. Blackwell.

Parikh, A., E. R. Miranda, M. Katoh-Kurasawa, D. Fuller, G. Rot, L. Zagar, T. Curk, et al. 2010. “Conserved Developmental Transcriptomes in Evolutionarily Divergent Species.” *Genome Biology* 11 (3): R35. https://doi.org/10.1186/gb-2010-11-3-r35.

Parkinson, K., P. Bolourani, D. Traynor, N. L. Aldren, R. R. Kay, G. Weeks, and C. R. L. Thompson. 2009. “Regulation of *Rap1* Activity Is Required for Differential Adhesion, Cell-Type Patterning and Morphogenesis in *Dictyostelium*.” *Journal of Cell Science* 122 (3): 335–44. https://doi.org/10.1242/jcs.036822.

Parkinson, K., N. J. Buttery, J. B. Wolf, and C. R. L. Thompson. 2011. “A Simple Mechanism for Complex Social Behavior.” *PLOS Biology* 9 (3): e1001039. https://doi.org/10.1371/journal.pbio.1001039.

Parsch, J., Z. Zhang, and J. F. Baines. 2009. “The Influence of Demography and Weak Selection on the McDonald-Kreitman Test: An Empirical Study in *Drosophila*.” *Molecular Biology and Evolution* 26 (3): 691–98. https://doi.org/10.1093/molbev/msn297.

Pertea, G. (2015) 2017. *Gffread: GFF/GTF Utility Providing Format Conversions, Region Filtering, FASTA Sequence Extraction and More*. C++. https://github.com/gpertea/gffread.

Pfeifer, B., U. Wittelsbürger, S. E. R. Onsins, and M. J. Lercher. 2014. "PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R." *Molecular Biology and Evolution*, April, msu136. https://doi.org/10.1093/molbev/msu136.

Pimentel, H., N. L. Bray, S. Puente, P. Melsted, and L. Pachter. 2017. "Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty." *Nature Methods* 14 (7): 687–90. https://doi.org/10.1038/nmeth.4324.

Purandare, S. R., R. D. Bickel, J. Jaquiery, C. Rispe, and J. A. Brisson. 2014. "Accelerated Evolution of Morph-Biased Genes in Pea Aphids." *Molecular Biology and Evolution* 31 (8): 2073–83. https://doi.org/10.1093/molbev/msu149.

Rice, W. R., and B. Holland. 1997. "The Enemies within: Intergenomic Conflict, Interlocus Contest Evolution (ICE), and the Intraspecific Red Queen." *Behavioral Ecology and Sociobiology* 41 (1): 1–10. https://doi.org/10.1007/s002650050357.

Robinson, G. E. 1999. "Integrative Animal Behaviour and Sociogenomics." *Trends in Ecology & Evolution* 14 (5): 202–5. https://doi.org/10.1016/S0169-5347(98)01536-5.

———. 2002. "Sociogenomics Takes Flight." *Science* 297 (5579): 204–5. https://doi.org/10.1126/science.1074493.

Robinson, G. E., R. D. Fernald, and D. F. Clayton. 2008. "Genes and Social Behavior." *Science (New York, N.Y.)* 322 (5903): 896–900. https://doi.org/10.1126/science.1159277.

Robinson, G. E., C. M. Grozinger, and C. W. Whitfield. 2005. "Sociogenomics: Social Life in Molecular Terms." *Nature Reviews Genetics* 6 (4): 257. https://doi.org/10.1038/nrg1575.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics (Oxford, England)* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, M. D., and A. Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3): R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Rocha, E. P. C., and E. J. Feil. 2010. "Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria?" *PLOS Genetics* 6 (9): e1001104. https://doi.org/10.1371/journal.pgen.1001104.

Rosengarten, R. David, B. Santhanam, D. Fuller, M. Katoh-Kurasawa, W. F. Loomis, B. Zupan, and G. Shaulsky. 2015. "Leaps and Lulls in the Developmental Transcriptome of *Dictyostelium discoideum*." *BMC Genomics* 16: 294. https://doi.org/10.1186/s12864-015-1491-7.

Santorelli, L. A., C. R. L. Thompson, E. Villegas, J. Svetz, C. Dinh, A. Parikh, R. Sucgang, et al. 2008. "Facultative Cheater Mutants Reveal the Genetic Complexity of Cooperation in Social Amoebae." *Nature* 451 (7182): 1107–10. https://doi.org/10.1038/nature06558.

Saxer, G., P. Havlak, S. A. Fox, M. A. Quance, S. Gupta, Y. Fofanov, J. E. Strassmann, and D. C. Queller. 2012. "Whole Genome Sequencing of Mutation Accumulation Lines Reveals a Low Mutation Rate in the Social Amoeba *Dictyostelium discoideum*." *PLoS ONE* 7 (10). https://doi.org/10.1371/journal.pone.0046759.

Sayres, M. A. Wilson. 2018. "Genetic Diversity on the Sex Chromosomes." *Genome Biology and Evolution* 10 (4): 1064–78. https://doi.org/10.1093/gbe/evy039.

Schupbach, T., and E. Wieschaus. 1986. "Germline Autonomy of Maternal-Effect Mutations Altering the Embryonic Body Pattern Of *Drosophila*." *Developmental Biology* 113 (2): 443–448.

Sedlazeck, F. J., P. Rescheneder, and A. von Haeseler. 2013. "NextGenMap: Fast and Accurate Read Mapping in Highly Polymorphic Genomes." *Bioinformatics (Oxford, England)* 29 (21): 2790–91. https://doi.org/10.1093/bioinformatics/btt468.

Settepani, V., J. Bechsgaard, and T. Bilde. 2016. "Phylogenetic Analysis Suggests That Sociality Is Associated with Reduced Effectiveness of Selection." *Ecology and Evolution* 6 (2): 469–77. https://doi.org/10.1002/ece3.1886.

Sharp, P. M., and K. M. Devine. 1989. "Codon Usage and Gene Expression Level in *Dictyostelium discoideum*: Highly Expressed Genes Do 'prefer' Optimal Codons." *Nucleic Acids Research* 17 (13): 5029–39.

Shaulsky, G., and R. H. Kessin. 2007. "The Cold War of the Social Amoebae." *Current Biology* 17 (16): R684–92. https://doi.org/10.1016/j.cub.2007.06.024.

Spradling, A. C. 1993. "Germline Cysts: Communes That Work." *Cell* 72 (5): 649–51. https://doi.org/10.1016/0092-8674(93)90393-5.

Stauber, M., H. Jackle, and U. Schmidt-Ott. 1999. "The Anterior Determinant *Bicoid* of *Drosophila* Is a Derived *Hox* Class 3 Gene." *Proceedings of the National Academy of Sciences of the United States of America* 96 (7): 3786–89.

Stauber, M., A. Prell, and U. Schmidt-Ott. 2002. "A Single *Hox3* Gene with Composite *Bicoid* and *Zerknullt* Expression Characteristics in Non-Cyclorrhaphan Flies." *Proceedings of the National Academy of Sciences of the United States of America* 99 (1): 274–79. https://doi.org/10.1073/pnas.012292899.

Stoletzki, N., and A. Eyre-Walker. 2011. "Estimation of the Neutrality Index." *Molecular Biology and Evolution* 28 (1): 63–70. https://doi.org/10.1093/molbev/msq249.

Strassmann, J. E., Y. Zhu, and D. C. Queller. 2000. "Altruism and Social Cheating in the Social Amoeba *Dictyostelium discoideum*." *Nature* 408 (6815): 965–67. https://doi.org/10.1038/35050087.

Sucgang, R., A. Kuo, X. Tian, W. Salerno, A. Parikh, C. L. Feasley, E. Dalin, et al. 2011. "Comparative Genomics of the Social Amoebae *Dictyostelium discoideum* and *Dictyostelium purpureum*." *Genome Biology* 12: R20. https://doi.org/10.1186/gb-2011-12-2-r20.

Sueoka, N. 1988. "Directional Mutation Pressure and Neutral Molecular Evolution." *Proceedings of the National Academy of Sciences* 85 (8): 2653–57. https://doi.org/10.1073/pnas.85.8.2653.

———. 1962. “On the Genetic Basis of Variation and Heterogeneity of Dna Base Composition.” *Proceedings of the National Academy of Sciences* 48 (4): 582–92. https://doi.org/10.1073/pnas.48.4.582.

Swire, J. 2007. “Selection on Synthesis Cost Affects Interprotein Amino Acid Usage in All Three Domains of Life.” *Journal of Molecular Evolution* 64 (5): 558–71. https://doi.org/10.1007/s00239-006-0206-8.

Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123 (3): 585–95.

Travisano, M., and G. J. Velicer. 2004. “Strategies of Microbial Cheater Control.” *Trends in Microbiology* 12 (2): 72–78. https://doi.org/10.1016/j.tim.2003.12.009.

Trotta, E. 2013. “Selection on Codon Bias in Yeast: A Transcriptional Hypothesis.” *Nucleic Acids Research* 41 (20): 9382–95. https://doi.org/10.1093/nar/gkt740.

Turner, P. E., and L. Chao. 1999. “Prisoner’s Dilemma in an RNA Virus.” *Nature* 398 (6726): 441–43. https://doi.org/10.1038/18913.

Urrutia, A. O., and L. D. Hurst. 2001. “Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection.” *Genetics* 159 (3): 1191–99.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, et al. 2013. “From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 43: 11.10.1-33. https://doi.org/10.1002/0471250953.bi1110s43.

Van Dyken, J. D., and M. J. Wade. 2010. “The Genetic Signature of Conditional Expression.” *Genetics* 184 (2): 557–70. https://doi.org/10.1534/genetics.109.110163.

———. 2012. “Detecting the Molecular Signature of Social Conflict: Theory and a Test with Bacterial Quorum Sensing Genes.” *The American Naturalist* 179 (4): 436–50. https://doi.org/10.1086/664609.

Vicoso, B., and B. Charlesworth. 2006. “Evolution on the X Chromosome: Unusual Patterns and Processes.” *Nature Reviews Genetics* 7 (8): 645. https://doi.org/10.1038/nrg1914.

Wagner, A. 2005. “Energy Constraints on the Evolution of Gene Expression.” *Molecular Biology and Evolution* 22 (6): 1365–74. https://doi.org/10.1093/molbev/msi126.

Wan, Y., K. Qu, Z. Ouyang, M. Kertesz, J. Li, R. Tibshirani, D. L. Makino, R. C. Nutter, E. Segal, and H. Y. Chang. 2012. “Genome-Wide Measurement of RNA Folding Energies.” *Molecular Cell* 48 (2): 169–81. https://doi.org/10.1016/j.molcel.2012.08.008.

Warner, M. R., A. S. Mikheyev, and T. A. Linksvayer. 2017. “Genomic Signature of Kin Selection in an Ant with Obligately Sterile Workers.” *Molecular Biology and Evolution* 34 (7): 1780–87. https://doi.org/10.1093/molbev/msx123.

West-Eberhard, M. J. 1979. “Sexual Selection, Social Competition, and Evolution.” *Proceedings of the American Philosophical Society* 123 (4): 222–34.

Wickham, H. 2016. *Ggplot2 - Elegant Graphics for Data Analysis*. http://www.springer.com/gb/book/9783319242750.

Wolf, J. B., E. D. Brodie III, J. M. Cheverud, A. J. Moore, and M. J. Wade. 1998. "Evolutionary Consequences of Indirect Genetic Effects." *Trends in Ecology & Evolution* 13 (2): 64–69. https://doi.org/10.1016/S0169-5347(97)01233-0.

Wolf, J. B., and M. J. Wade. 2009. "What Are Maternal Effects (and What Are They Not)?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1520): 1107–15. https://doi.org/10.1098/rstb.2008.0238.

Wolf, J. B., J. A. Howie, K. Parkinson, N. Gruenheit, D. Melo, D. Rozen, and C. R. L. Thompson. 2015. "Fitness Trade-Offs Result in the Illusion of Social Success." *Current Biology* 25 (8): 1086–90. https://doi.org/10.1016/j.cub.2015.02.061.

Woolley, S., J. Johnson, M. J. Smith, K. A. Crandall, and D. A. McClellan. 2003. "TreeSAAP: Selection on Amino Acid Properties Using Phylogenetic Trees." *Bioinformatics (Oxford, England)* 19 (5): 671–72.

Wright, S. 1931. "Evolution in Mendelian Populations." *Genetics* 16 (2): 97–159.

———. 1932. "The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution." In *Proceedings of the Sixth International Congress of Genetics*, 1:356–66.

Wysoker, A., K. Tibbetts, and T. Fennell. 2016. "Picard Tools." 2016. https://broadinstitute.github.io/picard/index.html.

Zhang, H., Y. Wang, J. Li, H. Chen, X. He, H. Zhang, H. Liang, and J. Lu. 2018. "Biosynthetic Energy Cost for Amino Acids Decreases in Cancer Evolution." *Nature Communications* 9 (1): 4124. https://doi.org/10.1038/s41467-018-06461-1.

Zuckerkandl, E., and L. Pauling. 1962. "Molecular Disease, Evolution, and Genetic Heterogeneity." In *Horizons in Biochemistry*, 189–225. New York: Academic Press.